Psychologisches Institut IV der Universität Düsseldorf

MÖGLICHKEITEN ZUR ERSTELLUNG VON ORDINALSKALEN

Habilitationsschrift zur Erlangung der Venia legendi im Fach Psychologie an der mathematisch naturwissenschaftlichen Fakultät der Universität Düsseldorf

vorgelegt von Christine Wolfrum

Düsseldorf, November 1983

UNIV.
LEIDEN
NED.

Psychologisches Institut IV der Universität Düsseldorf

MÖGLICHKEITEN ZUR ERSTELLUNG VON ORDINALSKALEN

Habilitationsschrift zur Erlangung der Venia legendi im Fach Psychologie an der mathematisch naturwissenschaftlichen Fakultät der Universität Düsseldorf

vorgelegt von Christine Wolfrum

Düsseldorf, November 1983

Inhaltsverzeichnis

A. Einleitung 1

B. Vorschläge zur Erstellung von Ordinalskalen 16

I. Ein Vorschlag zur Skalierung, wenn die Daten aus Rangreihen bestehen: Das ordinale Testmodell von Schulman 16

1. Das Modell der Klassischen Testtheorie übertragen auf Rangzahlen 20

1.1. Modell 20
1.2. Unabhängigkeit ordinaler Tests 24
1.2.1. Lokale Unkorreliertheit von Tests 24
1.2.2. Lineare experimentelle Unabhängigkeit von ordinalen Tests 25
1.2.3. Experimentelle Unabhängigkeit von ordinalen Tests 26
1.3. Gütemaße bei ordinalen Tests 27
1.3.1. Reliabilität von ordinalen Tests 27
1.3.2. Kriteriumsvalidität von Tests 27
1.4. Äquivalente Messungen 28
1.5. Paralleltests 30
1.6. Attenuationsformeln 32
1.7. Zusammensetzen von Tests 33
1.7.1. Zusammensetzen von Paralleltests und Veränderung der Reliabilität 33
1.7.2. Zusammensetzen lokal unkorrelierter Tests und Veränderung der Reliabilität 34
1.8. Veränderung der Validität beim Zusammensetzen von Tests 37
1.9. Korrelierte ordinale Tests 39
1.9.1. Reliabilität 40
1.9.2. Attenuation 41
1.9.3. Kriteriumsvalidität 42

2.	Schätzen von Parametern	45
2.1.	Schätzen des wahren Wertes eines festen Individuums	48
2.1.1.	Punktschätzung	48
2.1.1.1.	Die N Individuen der Population werden in eine Rangreihe gebracht	48
2.1.1.2.	Die n Individuen der Stichprobe werden in eine Rangreihe gebracht	48
2.1.2.	Konfidenzintervall für den wahren Wert	52
2.1.2.1.	Die N Individuen der Population werden in eine Rangreihe gebracht	52
2.1.2.2.	Die n Individuen der Stichprobe werden in eine Rangreihe gebracht	53
2.2.	Schätzen der Varianz der Rangzahlen des Individuums	56
2.2.1.	Die N Individuen der Population werden in eine Rangreihe gebracht	56
2.2.2.	Die n Individuen der Stichprobe werden in eine Rangreihe gebracht	57
2.3.	Schätzen der Reliabilität	63
2.3.1.	Die N Individuen der Population werden in eine Rangreihe gebracht	63
2.3.2.	Die n Individuen der Stichprobe werden in eine Rangreihe gebracht	66
2.4.	Schätzen der Korrelation zweier ordinaler Tests	71
2.4.1.	Die N Individuen der Population werden in eine Rangreihe gebracht	71
2.4.2.	Die n Individuen der Stichprobe werden in eine Rangreihe gebracht	71
2.5.	Zusammenfassung	73
3.	Anwendung des ordinalen Testmodells	75
3.1.	Skalieren von Items	75
3.1.1.	Vorgehen	75
3.1.2.	Itemanalyse	76
3.1.3.	Schätzen der Reliabilität und Validität der Rater	77
3.1.4.	Einordnen von Individuen auf dieser Skala	79

3.1.5.	Testprofile	79
3.1.6.	Konstruktion einer Leistungsskala	80
3.1.7.	Praktikabilität der beschriebenen Skalenkonstruktion	81
3.2.	Skalieren von Individuen	82
3.2.1.	Rater beurteilen Individuen	82
3.2.2.	Individuen beantworten Items	84
3.2.2.1.	Vorgehen	84
3.2.2.2.	Itemanalyse	85
3.2.2.3.	Score des Individuums	85
3.2.2.4.	Schätzen von Reliabilität und Validität	86
3.2.2.5.	Praktikabilität der beschriebenen Skalenkonstruktion	86
3.3.	Statistische Verfahren zum Vergleich von Rangreihen	87
3.3.1.	Vergleich zweier unabhängiger Gruppen von Ratern	87
3.3.1.1.	(2 × K!)-Felder-Test	88
3.3.1.2.	Vergleich zweier Anderson-Kannemann-Tafeln	90
3.3.1.3.	Vergleich zweier Friedman-Rangsummen-Vektoren	92
3.3.2.	Analyse einer Gruppe von Ratern	93
3.3.2.1.	Anpassungstest	94
3.3.2.2.	Anderson-Kannemann-Test	95
3.3.2.3.	Friedman-Test	95
3.4.	Beispiel	97
II.	Verwendung Log-linearer Modelle zur Skalierung	103
1.	Log-lineare Modelle	103
2.	Log-lineare Modelle und Skalierung	111
2.1.	Skalierung von Objekten (Items) mit einem speziellen Logit-Modell	114
2.2.	Äquivalenz zwischen probabilistischen Meßmodellen und Log-linearen Modellen	122
2.2.1.	Das Rasch Modell	122
2.2.2.	Das BTL-Modell	127
2.4.	Zusammenfassung	134
C.	Methodisches Resumée	135

Digitized by the Internet Archive
in 2026

https://archive.org/details/IA41543926_0018

A Einleitung

In der Arbeit werden im wesentlichen zwei Methoden zur Erstellung von Ordinal- bzw. Rangskalen untersucht. Hierbei handelt es sich einmal um die Übertragung der Konzepte der Klassischen Testtheorie auf Rangzahlen, zum anderen um ein spezielles Logit-Modell.

Die Meinungen darüber, ob die Beschäftigung mit Ordinalskalen in der Psychologie sinnvoll ist, sind unterschiedlich. Man kann z.B. die Auffassung vertreten, daß in den Abhandlungen zur modernen Meßtheorie wie z.B. in Suppes und Zinnes (1963) oder Pfanzagl (1968) alles Wesentliche über Messen auf Ordinalskalenniveau gesagt wurde und daß dem nichts hinzuzufügen sei. Aus der Sicht des Meßtheoretikers kann man dieser Ansicht zustimmen. Aus der Sicht eines Anwenders von Skalierungsverfahren jedoch kann man anderer Meinung sein. Häufig ist in Bereichen, wo z.B. Skalen ausgewählt und bezüglich ihrer Leistungsfähigkeit beurteilt werden müssen, ein Vergleich verschiedener Verfahren oder die Erstellung neuer Verfahren erforderlich.
Ist das betrachtete Merkmal nur auf Ordinalskalenniveau erfaßbar, so steht man vor der Aufgabe, ordinale Skalierungsverfahren auswählen oder neu konstruieren zu müssen. Ein Bedarf an ordinalen Skalierungsverfahren ist also vorhanden. Dieser wird noch deutlicher, wenn man bedenkt, daß fast alle Verfahren der eindimensionalen Skalierung, einschließlich der Klassischen Testtheorie, vorgeben, mindestens Intervallskalen zu erstellen. Man könnte sich fragen, ob es nicht eine Lösung wäre, diese Verfahren weiter zu benutzen und die entstehende Skala als Ordinalskala aufzufassen. Dieses Vorgehen kann jedoch nur ein Notbehelf für den Fall sein, daß keine ordinalen Skalierungsverfahren vorhanden sind, da z.B. bei der Konstruktion von Intervallskalen in der Regel Operationen durchgeführt werden, die für ordinales Datenniveau nicht zulässig sind. Um das zu vermeiden, sind Methoden zur Erstellung von Ordinalskalen erforderlich. Zum gegenwärtigen Zeitpunkt fehlt es auf der

einen Seite noch an derartigen Methoden, auf der anderen Seite werden Vorschläge zur Konstruktion von Ordinalskalen, wie z.B. das "ordinale Testmodell" von Schulman (1975) und das U-Modell von Irtel und Schmalhofer (1981/82) nicht genutzt.

Die Tatsache, daß es einerseits Merkmale gibt, die höchstens ordinalskalierbar sind, daß es andererseits aber noch an Methoden zur Konstruktion von Ordinalskalen fehlt, zeigt die Wichtigkeit der Forschung auf diesem Gebiet. Dabei sind nicht nur Neuentwicklungen wichtig und nützlich, sondern beispielsweise ebenso die Anpassung bekannter Verfahren an den ordinalen Fall, der Vergleich verschiedener Verfahren sowie die Beurteilung von Skalen bezüglich ihrer Leistungsfähigkeit.

Die beiden oben genannten Methoden, die Übertragung der Klassischen Testtheorie auf Rangzahlen und das spezielle Logit-Modell, stellen einen Beitrag zu diesem Gebiet dar. Es erscheint wichtig, diese beiden Verfahren nicht nur zu beschreiben, sondern sie auch bezüglich ihrer Meßeigenschaften beurteilen zu können. Aus diesem Grunde sollen hier, in der Einleitung, die wichtigsten Begriffe der Meßtheorie aufgeführt werden. Insbesondere erscheint der Unterschied zwischen fundamentaler und abgeleiteter Messung wichtig. Da sich der Begriff der abgeleiteten Messung im Verlaufe der Zeit wesentlich geändert hat, soll zunächst ein Blick auf die historische Entwicklung der Meßtheorie geworfen werden. Die Betrachtung einiger älterer Arbeiten, an denen man die Diskussion um das Messen in der Psychologie verfolgen kann, zeigt auch, daß die Frage, auf welchem Skalenniveau psychologische Merkmale gemessen werden können, vor 50 Jahren genauso aktuell war wie heute. Zwar hat sich der Begriff "Messen" geändert, die Probleme scheinen aber dieselben geblieben zu sein. Insbesondere wurde diskutiert, ob psychologische Merkmale auf einem höheren als ordinalen Skalenniveau meßbar seien. Obwohl die Diskussion um das ordinale Skalenniveau schon alt ist, also auch ein Bedarf an ordinalen Skalierungsverfahren bestand, wurden eher Verfahren zur Erstellung von Intervallskalen als von

Ordinalskalen entwickelt.

Die Betrachtung der historischen Entwicklung der Meßtheorie und der Schwierigkeiten des Messens in der Psychologie vermitteln einen Eindruck davon, welche der Probleme sich auch bei einer ordinalen Skalierung stellen werden.

Historisch entwickelte sich die Meßtheorie aus Untersuchungen zur Meßbarkeit physikalischer Eigenschaften (z.B. Helmholtz, 1887), aus wissenschaftstheoretischen Arbeiten zum Messen (z.B. Campbell, 1920) und auch aus der Nutzenmessung, welche sich bis zu Bernoulli zurückverfolgen läßt.

Das Konzept der "Klassischen Meßtheorie", wie es von Campbell vertreten wurde, hatte großen Einfluß auf die Diskussionen zur Frage der Meßbarkeit psychischer Eigenschaften. Campbell unterscheidet zwischen extensiven (oder additiven) und intensiven Merkmalen. Extensive Merkmale oder Quantitäten sind Merkmale, für die eine der Addition entsprechende empirische Operation existiert. Intensive Merkmale stellen Qualitäten dar. Hier existiert keine solche Operation. Weiterhin unterscheidet Campbell zwischen fundamentaler und abgeleiteter Messung. Abgeleitete Messung setzt die Messung fundamentaler Merkmale voraus. Als Beispiel hierfür wird stets die "Länge" als fundamental meßbar und die "Dichte" als abgeleitet meßbar angeführt. Nach Campbell sind nur extensive Merkmale fundamental meßbar. Diese Definition ist die Ursache für die Forderung der empirischen Interpretierbarkeit der Addition für eine fundamentale Meßbarkeit. Der enge Begriff der fundamentalen Messung nach Campbell hat zur Folge, daß es Messen auf Nominal- bzw. Ordinalskalenniveau nicht gibt.

Die strengen Forderungen der klassischen Meßtheorie wirkten sich auf die Psychologie so aus, daß diskutiert wurde, ob ein Messen in der Psychologie überhaupt möglich sei. Da Messen gleichgesetzt wurde mit Messen auf höherem als ordinalen Skalenniveau, wurde also diskutiert, welches Skalenniveau psychologische Variablen haben können. Die Diskussion betraf den gesamten Meßvorgang. Als Beispiel sei Scates (1937,a,b) angeführt. Für ihn war die wichtigste Forderung beim Meßvorgang die Entwicklung eines Konzeptes des zu messenden Merk-

mals. Der Psychologe müsse wissen, was er mißt. Häufig, so betont der Autor, sei die Variable, die gemessen wird, fern von dem Merkmal, welches ursprünglich gemessen werden sollte. Als Beispiel führt Scates die Messung der Intelligenz an. Auch bezüglich des Skalenniveaus äußert er sich kritisch. Seiner Meinung nach können viele Konzepte nur mit Größer-Kleiner-Vergleichen behandelt werden. Auch Johnson (1936) meint, daß Messen auf höherem als ordinalen Niveau kaum möglich sei. Er führt den Drang nach Quantifizierung darauf zurück, daß "quantitativ" mit "wissenschaftlich" gleichgesetzt würde. Aus diesem Grunde würde die Liste der gemessenen Merkmale auch größer mit jeder Ausgabe einer psychologischen Zeitschrift. McGregor (1935) wiederum setzt psychologisches Messen gleich physikalischem Messen. Durch die Methode der Operationalisierung verschwinde der Unterschied zwischen psychologischen und physikalischen Merkmalen, somit zwischen Psychologie und Physik. Gleichzeitig jedoch beklagt der Autor den Verlust an Psychologie in der Wissenschaft, den diese Methode zur Folge habe.
Weitzenhoffer (1951) schließlich kommt zu dem Schluß, daß in der Psychologie die Korrespondenz zwischen psychologischen Größen und Zahlen sehr begrenzt sei. Er empfiehlt, die Bemühungen, augenscheinlich unpassende Daten in existierende mathematische Strukturen zu pressen, aufzugeben.

Die hier angegebenen wenigen Artikel vermitteln den Eindruck, daß die Diskussion um die Meßbarkeit von Merkmalen vor 50 Jahren inhaltlicher geführt wurde als heute. Es wurde vorwiegend über die Art der Merkmale diskutiert und die damit verbundenen Probleme des Messens in der Psychologie. Das änderte sich, als sich die moderne Meßtheorie in den fünfziger Jahren des Problems annahm. Die Diskussionen beschäftigten sich jetzt eher mit den Voraussetzungen des Messens als mit dem zu messenden Merkmal. Die moderne Meßtheorie wurde aus der mathematischen Modelltheorie entwickelt. Die Probleme des Messens in der Psychologie wurden nicht berücksichtigt.

Trotz der Forderungen der klassischen Meßtheorie wurden in der Psychologie Konzepte entwickelt, welche diese Forderungen nicht beachteten. Es sind Versuche, den engen Rahmen der klassischen

Meßtheorie zu sprengen und spezielle Meßkonzepte für die Psychologie zu entwickeln. Es seien Thurstone (1927) mit dem Gesetz vom Vergleichsurteil, Coombs (1950) mit dem Unfolding angeführt. Auch die Verfahren der klassischen eindimensionalen Skalierung, wie z.B. die Methode der gleicherscheinenden Intervalle (Thurstone und Chave, 1929) und die Likert-Skalierung (Likert, 1932) entstanden. Zu nennen ist auch Stevens (1946) mit der Einteilung der Skalenniveaus in Nominal-, Ordinal-, Intervall- und Verhältnisskala.

Die Begründung der modernen Meßtheorie kann in den Arbeiten von Scott und Suppes (1958) und Suppes und Zinnes (1963) gesehen werden. Wichtige Beiträge sind weiterhin die Bücher von Pfanzagl (1968), Fishburn (1970), Krantz, Luce, Suppes und Tversky (1971). Die Grundbegriffe des Messens werden mit Hilfe der Grundbegriffe "Relativ" und "Repräsentation" der mathematischen Modelltheorie formuliert. Auch hier, wie in der klassischen Meßtheorie, wird zwischen fundamentaler und abgeleiteter Messung unterschieden. Zunächst soll auf die fundamentale Messung eingegangen werden. Man unterscheidet empirische und numerische Systeme, auch Relative oder Relationensysteme genannt. Ein Relativ wird mit $\mathfrak{A} = \langle A_1,\ldots,A_m;R_1,\ldots,R_n\rangle$ bezeichnet. Dabei sind $A_1,\ldots,A_m$ nichtleere Mengen von Objekten und $R_1,\ldots,R_n$ sind Relationen, welche auf den Mengen gelten. Ein Relativ muß mindestens eine Menge und mindestens eine Relation enthalten. Im folgenden wird der Einfachheit halber das Relativ mit nur einer Menge geschrieben.

Ein Relativ $\mathfrak{A} = \langle A,R_1,\ldots,R_n\rangle$ heißt empirisches Relativ, wenn A eine Menge empirischer Objekte ist und $R_1,\ldots,R_n$ beobachtbare empirische Relationen darstellen. Im Rahmen der Meßtheorie werden die empirischen Objekte vorausgesetzt. Über deren Herkunft oder deren Art wird nichts ausgesagt.
Ein Relativ $\mathfrak{A} = \langle B,S_1,\ldots,S_n\rangle$ heißt numerisches Relativ, wenn B eine Menge numerischer Objekte (Zahlen, Vektoren) ist und $S_1,\ldots,S_n$ Relationen auf dieser Menge bezeichnen.
Ein empirisches Relativ kann auf ein numerisches Relativ mit gleicher Anzahl von Relationen abgebildet werden, wenn die Relative

UNIV. LEIDEN

vom gleichen Typ sind. Dabei bedeutet "gleicher Typ", daß die Stelligkeiten der Relationen der beiden Relative übereinstimmen. Es wird definiert (Orth, 1974, S. 16):

"Es seien $\mathfrak{A} = \langle A, R_1, \ldots, R_n \rangle$ und $\mathfrak{B} = \langle B, S_1, \ldots, S_n \rangle$ zwei Relative desselben Typs. Eine Abbildung φ von A in B heißt homomorphe Abbildung (oder: Homomorphismus von A in B), wenn für alle Elemente $a_1, a_2 \in A$ und für alle $i=1,2,\ldots,n$ gilt $\varphi[R_i(a_1,a_2)] = S_i[\varphi(a_1),\varphi(a_2)]$.

Damit wird nicht nur die Menge A in die Menge B abgebildet, sondern es werden auch die Relationen abgebildet. Die Relationen R_i werden so auf die Relationen S_i des anderen Relativs abgebildet, daß das Bild einer Relation R_i zwischen zwei Elementen aus A die Relation S_i zwischen den Bildern dieser Elemente ist. Besteht ein Homomorphismus zwischen zwei Relativen $\mathfrak{A}$ und $\mathfrak{B}$, so wird $\mathfrak{A}$ durch $\mathfrak{B}$ "repräsentiert". Im Rahmen der Meßtheorie wird bei einer fundamentalen Messung ein empirisches Relativ in ein numerisches abgebildet. Das empirische Relativ wird durch das numerische Relativ repräsentiert. Die Existenz einer homomorphen Abbildung ist das Kriterium dafür, ob eine Zuordnung von Zahlen zu Objekten als "Messen" bezeichnet werden kann, ob also das betreffende Merkmal meßbar ist.

Das Repräsentationsproblem betrifft somit die Frage, ob Messung möglich ist. Im Repräsentationstheorem werden in Form von Axiomen diejenigen Eigenschaften angegeben, welche ein empirisches System $\mathfrak{A}$ zu erfüllen hat, damit es durch ein bestimmtes numerisches System $\mathfrak{B}$ repräsentiert werden kann.

Das Tripel $\langle \mathfrak{A}, \mathfrak{B}, \varphi \rangle$, wobei φ die homomorphe Abbildung darstellt, wird als "fundamentale Skala" bezeichnet.
Je nach Grad der Eindeutigkeit der homomorphen Abbildung werden verschiedene Skalentypen unterschieden. Statt von der "Meßbarkeit eines Merkmals" zu sprechen, spricht man genauer von der "Meßbarkeit auf einem bestimmten Skalenniveau". Das Eindeutigkeitsproblem befaßt sich mit der Frage der zulässigen Transformationen im Hinblick auf eine bestimmte Repräsentation. Speziell ist eine Ordinalskala eindeutig bis auf monotone Transformationen.
Neben dem Repräsentations- und Eindeutigkeitsproblem sei schließlich noch das dritte Hauptproblem der Meßtheorie erwähnt, das Bedeutsamkeitsproblem. Dieses stellt sich erst nach der Durch-

führung einer Messung. Für eine gegebene Skala ist eine numerische Aussage "bedeutsam" (sinnvoll, zulässig), wenn sich deren Wahrheit unter den zulässigen Skalentransformationen nicht ändert.

Die vorangehenden Ausführungen bezogen sich auf die Theorie der fundamentalen Messung. Während man von fundamentaler Messung nur sprechen kann, wenn ein empirisches Relativ homomorph in ein numerisches Relativ abgebildet wird, basiert die abgeleitete Messung nicht direkt auf einem empirischen relationalen System. Abgeleitete Messung setzt schon eine Zuweisung von Zahlen zu den empirischen Objekten voraus, also eine Messung.

Anstelle des empirischen Relativs wird ein abgeleitetes Meßsystem $\mathfrak{B}=\langle B,f_1,\ldots,f_n\rangle$ betrachtet. Hierbei ist B eine nichtleere Menge von Objekten und $f_1,\ldots,f_n$ sind Funktionen auf B. Sei g eine Funktion auf B und R eine Relation zwischen $f_1,\ldots,f_n$ und g. Das Tripel $\langle\mathfrak{B},R,g\rangle$ wird als abgeleitete Skala bezeichnet. R ist die repräsentierende Relation der Skala. Meistens handelt es sich hierbei um eine Gleichung. g stellt die abgeleitete Messung dar. Auch für die Theorie der abgeleiteten Messung ist das Repräsentations- und Eindeutigkeitsproblem von zentraler Bedeutung. Das Repräsentationsproblem wird gelöst, indem die Existenz der abgeleiteten numerischen Zuweisung g gezeigt wird. Für ein gegebenes abgeleitetes Meßsystem erhält man für unterschiedliche repräsentierende Relationen auch unterschiedliche abgeleitete Skalen, die verschiedenes Skalenniveau haben können.

Das Skalenniveau einer abgeleiteten Skala bezieht sich auf die zulässigen Transformationen von g. Das Skalenniveau der abgeleiteten Skala sagt nichts darüber aus, auf welchem Skalenniveau das Merkmal fundamental meßbar ist.
Es werden zwei Formen abgeleiteter Messung unterschieden: Statistiken und probabilistische Meßmodelle. Bei den probabilistischen Meßmodellen stellt die repräsentierende Relation eine prüfbare Hypothese dar, bei den Statistiken eine Definition (Gigerenzer, 1981, S. 121).

Die moderne Meßtheorie führte zu einer allgemeineren Definition der Messung als die klassische Meßtheorie. Die empirische Interpretation der Addition wird für eine Messung nicht mehr vorausgesetzt, sie ist jetzt ein Sonderfall (extensive Meßstruktur,

z.B. Krantz u.a., 1971). Genau wie die klassische Meßtheorie ist die moderne Meßtheorie jedoch nicht auf die speziellen Probleme des Messens in der Psychologie zugeschnitten, sondern paßt gleichermaßen auf die Physik wie auf die Psychologie. In den Meßtheorien werden die charakteristischen Probleme psychologischer Fragestellungen nicht berücksichtigt.

Fundamentale Kritik an der modernen Meßtheorie übt Gigerenzer (1981). Er kritisiert die Trennung zwischen psychologischer Theoriebildung und Messung, sowie das "Zuordnen von Zahlen zu Objekten", welches sowohl beim fundamentalen als auch beim abgeleiteten Messen vor sich geht. Für ihn ist Messen in der Psychologie viel mehr psychologische Modellbildung mit Hilfe numerischer Systeme. Gigerenzer knüpft an die Diskussion an, die vor der Entwicklung der Meßtheorien über das Messen in der Psychologie geführt wurde.
Er kritisiert die "Werkzeugfunktion" von Skalierungs- und Meßverfahren, d.h. die Forschungsstrategie, zunächst durch die Anwendung einer Skalierungsmethode die Merkmale zu skalieren und dann auf der Basis der berechneten Skalenwerte Theorien zu prüfen und zu bilden. Dieses Vorgehen bedeutet eine Trennung von Messung und Theoriebildung.
Er möchte Messung als theorieerzeugenden, d.h. modellbildenden Prozeß verstanden wissen. Aufbauend auf der modernen Meßtheorie und einer fünfstelligen Modellrelation entwickelt der Autor ein Konzept des Messens in der Psychologie. Der Begriff der Modellrelation sei näher erläutert: Die moderne Meßtheorie baut auf einer zweistelligen Modellrelation M(N,E) auf. N steht hier für ein numerisches Relativ, E für ein empirisches Relativ. Nehmen z.B. Individuen die Abbildung des empirischen in das numerische Relativ vor, wie es der Fall ist, wenn Rater Objekte oder Individuen beurteilen, so wird vorausgesetzt, daß die Rater sich wie objektive Meßinstrumente verhalten. Man wird statt dessen jedoch eher mit einer Interaktion zwischen Rater, zu beurteilendem Objekt und auch dem Merkmal rechnen müssen. Diesem trägt die fünfstellige Modellrelation $M(S,Z,N,E,\psi)$ Rechnung. Hier bedeutet S das forschende Individuum, Z das Ziel der Forschung, ψ den Gegenstandsbereich, über den Aussagen gemacht werden sollen.

Damit kann man sagen, daß Gigerenzer z.T. alte Forderungen neu stellt. Es sei an Scates (1937, a, b) erinnert, welcher forderte, daß der Psychologe ein "Konzept" haben müsse, bevor er anfängt zu messen. Die geforderte Einbettung des Messens in psychologische Theoriebildung klang schon in den älteren Arbeiten an.

Neben diesen fundamentalen Problemen stellt die "Natur" der zu messenden Merkmale eine weitere Schwierigkeit dar. Diese sind in der Regel komplexer Natur und häufig nicht direkt beobachtbar. Orth schreibt (1974, S. 19): "... hier liegen die Probleme des Messens noch vornehmlich im Auffinden psychologisch relevanter Eigenschaften, die fundamental meßbar sind." Damit gibt es kaum fundamentales Messen, und in den meisten Fällen kann nur eine abgeleitete Skala erstellt werden. Das Skalenniveau der abgeleiteten Skala ist abhängig von der repräsentierenden Relation. Bei der abgeleiteten Messung wird nichts über das Skalenniveau der Fundamentalskala ausgesagt. Wenn das Merkmal, welches gemessen werden soll, selbst nicht beobachtbar ist, wird statt dessen eine Indikatorvariable erfaßt. Man erhält damit keine Skala für das gewünschte Merkmal, sondern für die wirklich beobachtete Indikatorvariable.

Es können 2 Auffassungen vertreten werden:

1. Es existiert eine latente Variable, welche die Indikatorvariable, und damit die Daten, bedingt.
2. Eine latente Variable existiert nicht. Die Skalenwerte sind statistische Kenngrößen bezüglich der Indikatorvariablen.

Davison (1980) bezeichnet Skalierungsmethoden, welche mit diesen beiden Möglichkeiten arbeiten, als konfirmatorische bzw. explorative Methoden. Speziell für die in der Arbeit betrachtete ordinale Skalierung heißt das, daß im Fall 1 eine wahre Ordnung zwischen den zu skalierenden Objekten als existent angenommen wird. Diese bedingt die Daten. Die vom Skalierungsverfahren gelieferten Skalenwerte sind Schätzwerte für die wahren ordinalen Skalenwerte. Arbeitet das Skalierungsverfahren nicht mit einer zugrundeliegenden wahren Ordnung, so sind die Skalenwerte statistische Kenngrößen, welche eine Ordnung zwi-

schen den Objekten herstellen.

Wenn statt des Merkmals eine Indikatorvariable beobachtet wird, so hat man ein abgeleitetes Meßsystem $<B,f_1,\ldots,f_n>$, wobei sich die numerischen Zuordnungen $(f_1,\ldots,f_n)$ auf die Indikatorvariable beziehen. Die repräsentierende Relation gibt den Zusammenhang zwischen den $(f_1,\ldots,f_n)$ und den abgeleiteten Meßwerten an. Im Fall 1 werden die abgeleiteten Meßwerte so interpretiert, als wären sie Meßwerte auf der latenten Variablen. Dieses ist z.B. bei den probabilistischen Meßmodellen der Fall. Die repräsentierende Relation wird als Beziehung zwischen latenter Variable und Indikatorvariable aufgefaßt. Statt z.B. zu interpretieren: Die Indikatorvariable "Anzahl der mit "ja" beantworteten Items" wurde auf Absolutskalenniveau gemessen, wird behauptet, die latente Fähigkeit sei auf Absolutskalenniveau gemessen worden. Das Skalenniveau der Indikatorvariablen wird damit auf die latente Variable übertragen. Da weiterhin i.a. nicht zwischen abgeleiteter und fundamentaler Messung unterschieden wird, wird der Eindruck erweckt, als liege eine fundamentale Messung auf hohem Skalenniveau vor.
Im Fall 2 erfolgt die Interpretation bezüglich der wirklich beobachteten Indikatorvariablen.

Bei den meisten Meßvorgängen in der Psychologie werden nicht Merkmale fundamental gemessen, sondern es werden abgeleitete Skalen für Indikatorvariablen erstellt. Dabei hängt das Skalenniveau der abgeleiteten Skala von der repräsentierenden Relation ab. Suppes und Zinnes (1963) geben für ein BTL-System sieben repräsentierende Relationen an, die zu unterschiedlichen Niveaus der abgeleiteten Skala führen. Auch beim Thurstone-Modell handelt es sich um eine abgeleitete Messung. Häufig werden die abgeleiteten Meßsysteme repräsentierende Relationen haben, die nicht empirisch überprüfbar sind. Dieses ist z.B. der Fall bei der Klassischen Testtheorie und bei einigen Verfahren der klassischen eindimensionalen Skalierung.

Auch bei der Anwendung dieser Verfahren wird i.a. nicht zwischen abgeleiteter und fundamentaler Messung unterschieden, so daß man auch hier meinen könnte, es liege eine fundamentale Messung des Merkmals auf hohem Skalenniveau vor.
Die vorangegangenen Ausführungen über Meßtheorie zeigen jedoch, daß dieses keineswegs der Fall ist. Für die Fundamentalskala kann nicht ohne weiteres ein hohes Skalenniveau angenommen werden. Hierüber kann nur durch zusätzliche Annahmen etwas ausgesagt werden. Erstellen z.B. Rater die numerischen Zuweisungen $(f_1,\ldots,f_n)$ durch Beurteilung der Objekte bezüglich des zu messenden Merkmals, so kann das als Abbildung eines empirischen Relativs in ein numerisches Relativ aufgefaßt werden. Nimmt man an, daß die Rater ordinal urteilen, so handelt es sich um eine fundamentale Messung auf Ordinalskalenniveau. Verfahren, die für die Fundamentalskala eine Ordinalskala ergeben, können damit auch als ordinale Skalierungsverfahren bezeichnet werden.

Da die numerischen Zuweisungen $(f_1,\ldots,f_n)$ Grundlage für die abgeleitete Skala sind, soll schließlich deren Zustandekommen betrachtet werden.

Die Methoden der klassischen eindimensionalen Skalierung lassen sich einteilen in Subjekt-zentrierte Methoden, Reiz-zentrierte Methoden und Response-Methoden (Torgerson, 1958).
Mit Hilfe der Subjekt-zentrierten Methoden werden den Individuen Skalenwerte zugewiesen. Die Items, welche von den Individuen auf einer K-Punkte-Skala beantwortet werden, werden als Replikationen betrachtet. Ein Beispiel hierfür ist die Likert-Skalierung. Pro Individuum entsteht eine Häufigkeitsverteilung von Selbstbeurteilungen.
Mit Hilfe der reizzentrierten Methoden erhalten Items Skalenwerte zugewiesen. Rater beurteilen die zu skalierenden Items bezüglich ihrer Merkmalsausprägung auf einer K-Punkte-Skala. In diesem Falle stellen die Rater die Replikationen dar. Pro Item entsteht eine Häufigkeitsverteilung der Urteile über den Kategorien. Ein Beispiel hierfür ist die Methode der gleicherscheinenden Intervalle.

In beiden Fällen stellen die Häufigkeitsverteilungen die numerischen Zuweisungen dar. Jedem Individuum bzw. jedem Item i ist ein Vektor $\hat{p}_i = (\hat{p}_1,\ldots,\hat{p}_K)$ der relativen Häufigkeiten der Urteile zugeordnet.

Mit Hilfe der Response-Methoden sollen sowohl Individuen als auch Items skaliert werden. Beispiele hierfür sind das Rasch Modell und das U-Modell von Irtel und Schmalhofer. Handelt es sich um dichotome Items, so kann z.B. jedem Individuum die "Anzahl der mit ja beantworteten Items" zugeordnet werden.

Die beiden in der Arbeit betrachteten Methoden gehören zu den Reiz-zentrierten Verfahren.
Die zu skalierenden Objekte bzw. Individuen werden bezüglich ihrer Merkmalsausprägung von Ratern auf einer K-Punkte-Skala beurteilt bzw. in eine Rangreihe gebracht. Da es sich um Ordinalskalen handelt, müssen die Rater ordinal urteilen können. Sie müssen Präferenzurteile abgeben und transitiv urteilen können. Wenn also für die Objekte O_i, O_j und O_k gilt $O_i \succ O_j$ und $O_j \succ O_k$, dann muß auch $O_i \succ O_k$ gelten. Die Beurteilung der Objekte auf einer geordneten K-Punkte-Skala bzw. das Bilden einer Rangreihe stellt eine fundamentale Messung des beurteilten Merkmals auf Ordinalskalenniveau dar. Damit müssen die Rater beim Urteilen die Axiome der Meßtheorie zur Erstellung einer Ordinalskala erfüllen. Da man annehmen kann, daß es für Rater leichter ist, ordinal als z.B. auf Intervallskalenniveau zu urteilen, soll davon ausgegangen werden, daß, eventuell mit Hilfe eines Ratertrainings, die Axiome erfüllt werden. Von daher sind ordinale Skalierungsverfahren zu bevorzugen.

Beim ordinalen Testmodell von Schulman (1975, 1976, 1978), welches die Übertragung der Klassischen Testtheorie auf Rangzahlen bedeutet, bringt ein Rater N Individuen in eine Rangreihe. R_i sei die Rangzahl des Individuums i, i=1,...,N. Er beurteilt nach einer Wahrscheinlichkeitsverteilung für die N! möglichen Rangreihen die Individuen bezüglich ihrer Ausprägung auf einem bestimmten Merkmal. Aus der Wahrscheinlichkeitsverteilung des Raters für die Rangreihen kann für jedes Individuum eine Verteilung der Rangzahlen abgeleitet werden.

Damit wird jedem Individuum i, i=1,...,N eine Wahrscheinlichkeitsverteilung zugeordnet $p_i = (p_{i1},...,p_{iN})$. Ist B die Menge der Individuen, so stellt $\mathfrak{B} = \langle B, p_1,...,p_N \rangle$ ein abgeleitetes Meßsystem dar.

Durch die Definition

$$t_i = E[R_i], \quad i=1,...,N,$$

wobei $E[R_i]$ der Erwartungwert der Rangzahlen ist, wird jedem Individuum ein abgeleiteter Meßwert zugeordnet. t_i ist der wahre Wert. $t_i = E[R_i]$ ist als repräsentierende Relation empirisch nicht überprüfbar. Die Skalenwerte der Individuen t_i, i=1,...,N, liegen auf einer abgeleiteten Ordinalskala.
Die Art der Zuweisung von Zahlen zu Objekten kann damit als abgeleitete Messung bezeichnet werden mit einer repräsentierenden Relation, die aus einer Definition besteht.
Der Rater soll wie ein Meßinstrument funktionieren, d.h. objektiv urteilen. Eine Interaktion zwischen Rater und Individuum wird nicht berücksichtigt. Der Rater nimmt die fundamentale Messung des beobachteten Merkmals vor.

Im Abschnitt I der vorliegenden Arbeit wird das Modell ausführlich vorgestellt. Für alle wesentlichen Konzepte der Klassischen Testtheorie wird nachgewiesen, daß sie auch für Rangzahlen gültig sind. Anschließend werden erwartungstreue Schätzer für den wahren Wert eines Individuums, die Varianz der Rangzahlen des Individuums und die Reliabilität abgeleitet. Dabei werden jeweils die beiden Fälle betrachtet, daß die Population der Individuen zur Verfügung steht oder daß nur eine Stichprobe von Individuen vorhanden ist.

In Abschnitt II wird untersucht, ob Log-lineare Modelle als Skalierungs- bzw. Meßmodelle verwendet werden können. Log-lineare Modelle werden zur Analyse des Zusammenhanges von Variablen in Kontingenztafeln verwendet. Mit Hilfe dieser

Modelle wurde eine vollständige Assoziationsanalyse in mehrdimensionalen Kontingenztafeln möglich. Für den Logarithmus der Zellwahrscheinlichkeiten oder der Erwartungswerte der Zellhäufigkeiten wird ein lineares Modell angenommen. (Log-lineares Modell).
Über die Variablen, welche der Kontingenztafel zu Grunde liegen, werden keine Annahmen gemacht. Es werden i.a. keine latenten Variablen angenommen. In der Literatur der letzten Jahre findet man Hinweise auf Äquivalenzen zwischen psychometrischen Skalierungsmodellen und Log-linearen Modellen. Da es nur endlich viele Möglichkeiten gibt, leicht handhabbare Modellgleichungen zu formulieren, ist es zunächst nicht verwunderlich, wenn man ähnliche oder dieselben Modellgleichungen in verschiedenen Anwendungsgebieten findet. Es handelt sich zunächst um eine formale Äquivalenz. Die Methoden der Parameterschätzung können unterschiedlich sein und erst recht die Interpretation der Parameter.
Da bei psychometrischen Skalierungsmodellen i.a. latente Variablen als existent angenommen werden und die Daten sich nicht ohne weiteres mit einer Kontingenztafel darstellen lassen, kann von einer Äquivalenz von Skalierungsmodellen und Log-linearen Modellen nicht gesprochen werden. Wenn sich die Skalierungsdaten in einer Kontingenztafel darstellen lassen, kann man jedoch untersuchen, ob ein Log-lineares Modell als Skalierungs- bzw. Meßmodell eingesetzt werden kann.

Für eine Situation, die gleich derjenigen bei der Methode der gleicherscheinenden Intervalle ist, nämlich g Rater beurteilen R Items (bzw. Objekte) auf einer geordneten K-Punkte-Skala, wird ein spezielles Logit-Modell bezüglich seiner Anwendbarkeit als Skalierungsmodell untersucht. Das Modell erweist sich als abgeleitetes Meßmodell mit einer empirisch überprüfbaren repräsentierenden Relation. Es erstellt eine stochastische Ordnung der Items/Objekte. Die numerischen Zuweisungen für das abgeleitete Meßsystem entstehen auch hier durch Raterurteile. Die Interpretation der Parameter erfolgt bezüglich des von den Ratern beobachteten Merkmals. Das Modell kann als probabilistische Version der Methode der gleicherscheinenden Intervalle angesehen werden. Ersetzt man die

Rater durch Items, die Replikationen darstellen, und die Objekte durch Personen, so könnte man dieses als probabilistische Version der Likert-Skalierung ansehen. Das Verfahren wird zum Thema "Ordinalskalen" aufgeführt, da die Fundamentalskala höchstens ordinales Skalenniveau hat.

Log-lineare Repräsentationen für das Rasch Modell und das Bradley-Terry-Modell geben Mellenbergh und Vijn (1981), Kelderman (1984), Fienberg und Larntz (1976), Fienberg (1977), Fienberg (1979), Semenya u.a. (1983) und Beaver (1983). Eine Skalierung von Antwortmustern, die durch Beantwortung einer Guttman-Skala entstehen, mit Hilfe Log-linearer Modelle geben Fienberg (1977, S. 155), Goodman (1975) und Davison (1980).
Um den Zusammenhang von probabilistischen Skalierungsmodellen und Log-linearen Modellen durchsichtiger zu machen, werden das Rasch Modell und das Bradley-Terry-Modell wie Mellenbergh und Vijn (1981) bzw. Fienberg und Larntz (1976) es darstellen, untersucht.

Bezüglich der Gemeinsamkeiten und Unterschiede der beiden Verfahren kann abschließend gesagt werden: Beide Modelle sind abgeleitete Meßsysteme. Beim ordinalen Testmodell ist die repräsentierende Relation empirisch nicht überprüfbar, beim speziellen Logit-Modell ist sie empirisch überprüfbar. Sowohl beim ordinalen Testmodell als auch bei dem Logit-Modell entstehen die numerischen Zuordnungen für das abgeleitete Meßsystem durch Raterurteile. Von den Ratern wird ein objektives Urteil gefordert. Interaktionen zwischen den Ratern und den zu skalierenden Objekten werden in beiden Modellen nicht berücksichtigt. Die Anwendung des Logit-Modells ist durch die Anpassung des Modells an die Daten und die Schätzung der Parameter aufwendiger als die Anwendung des ordinalen Testmodells. Beide Verfahren sollten, wie Gigerenzer (1981) es fordert, in ihrer modellbildenden Funktion gesehen und entsprechend eingesetzt werden.

B. Vorschläge zur Erstellung von Ordinalskalen

I. Ein Vorschlag zur Skalierung, wenn die Daten aus Rangreihen bestehen: Das ordinale Testmodell von Schulman

Schulman und Haden (1975) übertrugen das Modell der Klassischen Testtheorie (KTT):
Beobachtungswert = wahrer Wert + Meßfehler
auf den Fall, daß einem Individuum bezüglich seiner Ausprägung auf einem bestimmten Merkmal eine Rangzahl zugewiesen wird. Diese Rangzahl stellt den Beobachtungswert dar. Sie ist nur sinnvoll in Bezug auf eine Population. Die N Individuen der Population werden bezüglich ihrer Merkmalsausprägungen auf dem betrachteten Merkmal in eine Rangreihe gebracht. Dieses kann beispielsweise auf Grund vorhandener Scores geschehen oder mit Hilfe eines (bzw. mehrerer) Raters, welcher die Individuen direkt beurteilt. Wenn man annimmt, daß die entstehende Rangreihe keine Bindungen aufweist, so erhält man die Rangzahlen 1,...,N.

Könnte der urteilende Rater die Individuen mehrere Male unabhängig voneinander in eine Rangreihe bringen, so entstünde durch Urteilsschwankungen nicht jedesmal dieselbe Rangreihe. Der Rater urteilt nach einer Wahrscheinlichkeitsverteilung für die N! möglichen Rangreihen. Aus der Wahrscheinlichkeitsverteilung für die N! Rangreihen kann für jedes Individuum die Verteilung der Rangzahlen abgeleitet werden. Es handelt sich hierbei um eine diskrete Verteilung, deren Form unbekannt ist.

Schulman und Haden (1975) bezeichnen einen Test, welcher vollständig bestimmt ist durch die Population der N Individuen und durch die Wahrscheinlichkeitsverteilung für die N! Rangreihen als "ordinalen Test". Die Bezeichnung "ordinaler Test", sowie dessen Abkürzung "O" sollen auch hier beibehalten werden. Bringt ein Rater die Individuen durch Beurteilung ihrer Merkmalsausprägung in eine Rangreihe, wobei er mit einer gewissen Wahrscheinlichkeit diese Rangreihe aus den N! möglichen Rangreihen auswählt, so repräsentiert in diesem Falle der Rater den ordinalen Test.

Zwei ordinale Tests können damit repräsentiert werden durch zwei Rater mit z.B. unterschiedlichen Wahrscheinlichkeitsverteilungen für die N! Rangreihen. Bei dieser Interpretation eines ordinalen Tests gelten die Konzepte der KTT für die Rater. So bedeutet die Unabhängigkeit zweier Tests, daß die beiden Rater beim Beurteilen der Individuen unabhängig voneinander vorgehen. Ebenso beziehen sich Reliabilität und Validität von Tests auf die Rater. Die Reliabilität eines Tests bedeutet die Reliabilität des Raters. Zur Bestimmung der Validität beurteilt ein Rater die Individuen bezüglich des betrachteten Merkmals, ein anderer Rater dieselben Individuen bezüglich einer Kriteriumsvariablen. Dieses Vorgehen ist insofern problematisch, als die Validität sehr stark von Unterschieden zwischen den Ratern beeinflußt werden kann. Ein Ausweg wäre, wenn derselbe Rater die Individuen bezüglich des Merkmals und auch bezüglich des Kriteriums beurteilt. Dann können sich jedoch korrelierte Fehler ergeben. Ein solches Vorgehen ist nur sinnvoll, wenn der Rater hochreliabel ist. Williams und Zimmerman (1982) und Zimmerman und Williams (1977) betrachteten dieses Problem im Rahmen der KTT. Im vorliegenden Kapitel wird für einige dieser Ergebnisse gezeigt, daß sie auch für Rangzahlen gelten.

Repräsentieren Rater die ordinalen Tests, so müssen diese wie "objektive" Meßinstrumente funktionieren. Im Rahmen des ordinalen Testmodells können die Rater unterschiedliche Funktionen ausüben

1. Die Rater stellen selbst ordinale Tests dar, mit möglicherweise unterschiedlichen Wahrscheinlichkeitsverteilungen,
2. für das Schätzen stellen die Rater die unabhängigen Wiederholungen eines ordinalen Tests dar. Es sind "homogene" Rater erforderlich, von denen jeder einem Individuum denselben wahren Wert und dieselbe Varianz der Rangzahlen zuweist.

Die Übertragung der KTT auf Rangzahlen bedeutet, daß analog zur KTT jetzt eine Rangzahl in einen wahren Wert und einen Meßfehler aufgespalten wird. Man kann unterschiedlicher

Meinung darüber sein, ob es sinnvoll ist, eine Rangzahl aufzuspalten. Für eine Übertragung der KTT auf Rangzahlen ist dieser Ansatz wesentlich. "Übertragen der KTT auf Rangzahlen" bedeutet, daß die Konzepte der KTT für die neue Situation auf ihre Gültigkeit hin überprüft werden müssen.

Für den vorliegenden Abschnitt I wurden Teile aus den Veröffentlichungen Schulmans (1975, 1976, 1978) verwendet, sowie Teile neu erstellt.

So wurden bei der Übertragung des Modells der Klassischen Testtheorie auf Rangzahlen die Unabhängigkeit von Tests, Kriteriumsvalidität von Tests, Einfluß der Testlänge auf die Validität von Tests, korrelierte Tests hinzugenommen. Beim Schätzen von Parametern (I.2.) wurden zu den Punkten 2.1., 2.2., 2.3. Teile neu erstellt (Schätzen bei bekannter Population, Konfidenzintervall für den wahren Wert, Schätzen bei unbekanntem Populationsumfang, Beweise beim Schätzen der Varianz und der Reliabilität). Die Punkte 3.1., 3.2. wurden neu erstellt. Bei Punkt 3.3. wurden die angegebenen bekannten statistischen Verfahren zur Analyse von Rangreihen verwendet.

1. Das Modell der Klassischen Testtheorie übertragen auf Rangzahlen

1.1. Modell

Es wird angenommen, daß eine endliche Population, bestehend aus N Individuen, gegeben ist. Ein Rater ordnet die Individuen bezüglich ihrer Merkmalsausprägungen in einer Rangreihe an. Sei R_i die Zufallsvariable "Rangplatz des Individuums i", so hat man bei N Individuen die Zufallsvariablen $R_1,\ldots,R_N$.

Der Rater weist den Individuen die Rangzahlen $j_1,\ldots,j_N$ mit der Wahrscheinlichkeit

$$P(\{R_1,\ldots,R_N\} = \{j_1,\ldots,j_N\}) = p_{j_1,\ldots,j_N}$$

zu.

Analog zur KTT wird für das Individuum i, i=1,...,N, ein *wahrer Wert* t_i definiert:

$$t_i = E[R_i] .$$

Der Erwartungswert wird über die Verteilung der Rangzahlen des Individuums i gebildet. Diese Verteilung erhält man aus der Wahrscheinlichkeitsverteilung des Raters für die N! möglichen Rangordnungen. Da sich die Bildung des Erwartungswertes auf die Verteilung der Ranganordnungen bezieht, wird die Bezeichnung "E_a" eingeführt. Der Index "a" soll auf "Anordnung" hinweisen.

Damit ist

$$t_i = E_a[R_i] = \sum_{j_1,\ldots,j_N} j_i p_{j_1,\ldots,j_N} .$$

"$\sum_{j_1,\ldots,j_N}$" bedeutet, daß über die N! Rangordnungen summiert wird, j_i ist die Rangzahl des i-ten Individuums in der Rangordnung $j_1,\ldots,j_N$.

Die Differenz zwischen der Rangzahl des Individuums i und seinem wahren Wert wird als *Meßfehler* definiert:

$$E_i = R_i - t_i = R_i - E_a[R_i] .$$

Hiermit und mit der Definition des wahren Wertes als Erwartungswert erhält man für das Individuum i das Modell

$$R_i = t_i + E_i ,$$

welches demjenigen der KTT entspricht. Analog zur KTT ergeben sich als Folgerungen aus dem Modell:

1. Der Erwartungswert des Meßfehlers bezüglich der Rangordnungen ist gleich Null für ein festes Individuum:

   $E_a[E_i] = 0 .$

   B. $E_a[E_i] = E_a[R_i - t_i] = E_a[R_i] - E_a[t_i] = t_i - t_i = 0 .$

2. Der wahre Wert und der Meßfehler sind unkorreliert.

   $Cov_a[t_i, E_i] = 0 .$

   B. $Cov_a[t_i, E_i] = 0$, da t_i konstant ist.

Die bisherigen Überlegungen setzten ein festes Individuum i voraus. Nimmt man statt dessen an, daß aus der Population der N Individuen eines zufällig ausgewählt wird, so hat man in diesem Modell die Zufallsvariablen $(R_1,\ldots,R_N, I)$ zu betrachten. Dabei wird die Zufallsvariable I = "Auswahl eines Individuums" mit $P(I=i) = \frac{1}{N}$, $i=1,\ldots,N$, als stochastisch unabhängig von den Zufallsvariablen $(R_1,\ldots,R_N)$ vorausgesetzt. D.h., die Zuweisung der Rangzahlen durch den Rater hat unabhängig von der Auswahl des Individuums zu geschehen.

Erwartungswerte können sich jetzt auf den gemeinsamen Zufallsprozeß "Zuweisung der Rangzahlen und Auswahl des Individuums", wie auch auf die einzelnen Prozesse beziehen. Der Erwartungswert bezüglich der Zufallsauswahl des Individuums sei mit "E_i" bezeichnet, der Erwartungswert bezüglich des gemeinsamen Zufallsprozesses (Gesamterwartungswert) mit "E_g". Die Bezeichnung "E_g" wird auch benutzt, wenn es sich statt um die Verteilung eines Raters um die bivariate Verteilung von 2 Ratern handelt.

Für ein zufällig ausgewähltes Individuum werde analog zur KTT formuliert

$$R = T + E ,$$

wobei R die Rangzahl, T der wahre Wert und E der Meßfehler des zufällig ausgewählten Individuums sind. R,T und E sind Zufallsvariablen. Es ist

$$t = E_g[R] = E_i[E_a[R|I]] = \frac{1}{N} \sum_{i=1}^{N} \Big(\sum_{j_1,\dots,j_N} j_i p_{j_1,\dots,j_N} \Big) =$$

$$= \frac{1}{N} \sum_{j_1,\dots,j_N} p_{j_1,\dots,j_N} \sum_{i=1}^{N} j_i = \frac{1}{N} \sum_{j_1,\dots,j_N} p_{j_1,\dots,j_N} \cdot \frac{N(N+1)}{2} = \frac{N+1}{2} .$$

Mit $R = T + E$ erhält man nach Bildung des Erwartungswertes

$$E_g[R] = E_g[T] + E_g[E] .$$

Da der Erwartungswert des Meßfehlers gleich Null ist

$$E_g[E] = E_i[E_a[E|I]] = 0 ,$$

gilt

$$E_g[R] = E_g[T] = \frac{N+1}{2} .$$

An dieser Stelle sei erwähnt, daß für ein zufällig ausgewähltes Individuum der Meßfehler nicht als Differenz zwischen der Zufallsvariablen R und ihrem Erwartungswert $E[R]$ definiert ist, sondern als Differenz der Zufallsvariablen R und T: $E = R - T$. Wird das Individuum i ausgewählt, so gilt jedoch $E_i = R_i - t_i = R_i - E_a[R_i]$.

Es folgt aus $R_i = t_i + E_i$ durch Summation über i

$$\sum_{i=1}^{N} R_i = \sum_{i=1}^{N} t_i + \sum_{i=1}^{N} E_i .$$

Mit $\sum_{i=1}^{N} R_i = \frac{N(N+1)}{2}$ und $\sum_{i=1}^{N} t_i = N E_g[T] = \frac{N(N+1)}{2}$

erhält man

$$\frac{N(N+1)}{2} = \frac{N(N+1)}{2} + \sum_{i=1}^{N} E_i$$

d.h.

$$\sum_{i=1}^{N} E_i = 0 .$$

Für die Varianz der Rangzahlen erhält man

$$\mathrm{Var}_g[R] = \frac{N^2-1}{12} ,$$

wegen

$$\mathrm{Var}_g[R] = E_g[R^2] - (E_g[R])^2 = E_i[E_a[R^2|I]] - (\tfrac{N+1}{2})^2 =$$

$$= \frac{1}{N} \sum_{i=1}^{N} \Big(\sum_{j_1,\dots,j_N} j_i^2 \, p_{j_1,\dots,j_N} \Big) - (\tfrac{N+1}{2})^2 =$$

$$= \frac{1}{N} \sum_{j_1,\dots,j_N} p_{j_1,\dots,j_N} \cdot \sum_{i=1}^{N} j_i^2 - (\tfrac{N+1}{2})^2 = \frac{N^2-1}{12} .$$

$\mathrm{Var}_g[R]$ ist im Unterschied zur KTT hier eine Konstante. Man erhält dieses Ergebnis jedoch nur, wenn, wie es hier vorausgesetzt wurde, keine Bindungen vorliegen. Analog zur KTT ergeben sich als Folgerungen:

1. Bei Zufallsauswahl eines Individuums ist der Erwartungswert des Meßfehlers Null:

   $E_g[E] = 0.$

   B. $E_g[E] = E_i[E_a[E|I]] = 0.$

2. Bei Zufallsauswahl eines Individuums sind der wahre Wert und der Meßfehler unkorreliert:

   $\mathrm{Cov}_g[T,E] = 0.$

   B. $\mathrm{Cov}_g[T,E] = E_g[TE] - E_g[T]E_g[E] = E_g[TE] = E_i[E_a[TE|I]] =$

   $= E_i[t_i E_a[E|I]] = 0.$

Durch die Definition eines wahren Wertes t_i als Erwartungswert der Rangzahlen für das Individuum i, i=1,...,N, ergibt sich ein zur KTT analoges Modell, nämlich die Aufspaltung der Rangzahl in einen wahren Wert und einen Meßfehler.
Unterschiedlich zur KTT ergibt sich für ein zufällig ausgewähltes Individuum $E_g[T] = E_g[R] = (N+1)/2$ und $\mathrm{Var}_g[R] = (N^2-1)/12$. Ein weiterer Unterschied besteht darin, daß es sich hier um eine Zuweisung von Rangzahlen bei einer endlichen Population handelt.
Es ist zu vermuten, daß durch die Definition des wahren Wertes als Erwartungswert die Konzepte der KTT übernommen werden können. Wie

schon bemerkt wurde, ist jedoch ein Nachprüfen erforderlich. Schulmans Definitionen weichen teilweise von den entsprechenden der KTT ab, was jedesmal vermerkt wird.

1.2. Unabhängigkeit ordinaler Tests

In der KTT unterscheidet man die lokale Unkorreliertheit, die lineare experimentelle Unabhängigkeit und die stochastische Unabhängigkeit zweier Tests. Schulman und Hader (1975) definieren hiervon nur die lineare experimentelle Unabhängigkeit.

Im folgenden werden alle drei Definitionen angegeben. Werden die ordinalen Tests von 2 Ratern, welche die Individuen beurteilen, repräsentiert, so ist es schwer vorstellbar, daß diese zwar lokal unkorreliert, aber nicht stochastisch unabhängig voneinander die Individuen beurteilen. In solchen Situationen wird nur die stochastische Unabhängigkeit von Bedeutung sein.

1.2.1. Lokale Unkorreliertheit von Tests

Es werden 2 ordinale Tests betrachtet, welche dem Individuum i die Rangzahlen R_i bzw. R_i' zuordnen. Es wird angenommen, daß die Tests durch 2 Rater repräsentiert werden, welche die N Individuen beurteilen. Die Wahrscheinlichkeitsverteilungen der Rater für die N! möglichen Rangreihen können unterschiedlich sein. Analog zur KTT werden zwei derartige Tests als *lokal unkorreliert* bezeichnet, wenn für alle i gilt

$$Cov_{aa'}[R_i, R_i'] = 0 .$$

Dabei bedeutet $Cov_{aa'}$ die Bildung der Kovarianz bezüglich der bivariaten Verteilung der beiden Rater. Die Unkorreliertheit sagt etwas darüber aus, wie die Beurteilung durch die Rater vor sich gehen soll.

Als Folgerungen ergeben sich:

1. Betrachtet man statt des festen Individuums i eines, welches unabhängig von den Beurteilungen der beiden Rater mit der Wahrscheinlichkeit $\frac{1}{N}$ ausgewählt wurde, so gilt für zwei lokal unkorreliert urteilende Rater (zwei lokal unkorrelierte ordinale Tests):

   $$Cov_g[E, E'] = 0 .$$

B. $Cov_g[E,E'] = Cov_g[R-T,R'-T'] = E_g[(R-T)(R'-T')] -$

$$- E_g[(R-T)]E_g[(R'-T')] =$$

$$= E_g[(R-T)(R'-T')] = E_g[RR'-RT'-R'T+TT']$$

$$= E_g[TT'] + E_g[TE'] + E_g[ET'] + E_g[EE'] - E_g[TT'] -$$

$$- E_g[ET'] - E_g[TT'] - E_g[E'T] + E_g[TT']$$

$$= E_g[EE'] = E_i[E_{aa'}[EE'|I]] = E_i[E_{aa'}[E_iE_i']] .$$

$E_{aa'}[E_iE_i'] = E_a[E_{a'}[E_iE_i'|R]] = E_a[E_iE_{a'}[E_i'|R]] = 0$. Damit ist

$Cov_g[E,E'] = 0$,

d.h., die Meßfehler zweier lokal unkorrelierter ordinaler Tests sind unkorreliert.

2. Der wahre Wert des einen Tests und der Meßfehler des anderen Tests sind für zwei lokal unkorrelierte Tests unkorreliert.

$Cov_g[T,E'] = 0$, $Cov_g[T',E] = 0$.

B. $Cov_g[T,E'] = E_g[TE'] - E_g[T]E_g[E'] = E_g[TE'] =$

$$= E_i[E_{aa'}[TE'|I]] = E_i[t_iE_{aa'}[E'|I]] = 0 .$$

3. Für ein zufällig ausgewähltes Individuum ist für zwei lokal unkorrelierte Tests die Kovarianz der Rangzahlen gleich der Kovarianz der wahren Werte:

$Cov_g[R,R'] = Cov_g[T,T']$.

B. $Cov_g[R,R'] = Cov_g[T+E,T'+E'] = Cov_g[T,T'] + Cov_g[E,E'] + Cov_g[E,T'] +$

$$+ Cov_g[E',T] = Cov_g[T,T'] .$$

1.2.2. Lineare experimentelle Unabhängigkeit von ordinalen Tests

Schulman und Haden (1975) bezeichnen zwei Tests, welche die Folgerungen 1. und 2. aus der lokalen Unkorreliertheit erfüllen, als linear experimentell unabhängig. Zwei Tests werden auch *linear experimentell unabhängig* genannt, wenn

$E_a[R_i|R_i' = j_i'] = E_a[R_i]$ und

$E_{a'}[R_i'|R_i = j_i] = E_{a'}[R_i']$ gilt, i=1,...,N.

Betrachtet man

$$E_a[R_i | R'_i = j'_i] = E_a[R_i] ,$$

so bedeutet das: Nachdem Rater Nr. 2 den N Individuen die Rangzahlen $j'_1,\ldots,j'_N$ zugewiesen hat, bringt auch Rater Nr. 1, welcher die Rangordnung des Raters Nr. 2 kennt, die N Individuen in eine Rangreihe $j_1,\ldots,j_N$. Der wahre Wert des Individuums i, $t_i = E_a[R_i]$, der entsteht, wenn Rater Nr. 1 die Beurteilung des Raters Nr. 2 nicht kennt, soll gleich dem wahren Wert sein, der entsteht, wenn Rater Nr. 1 das Urteil des Raters Nr. 2 kennt. D.h., die Kenntnis des Urteils des Raters Nr. 2 soll sich bei Rater Nr. 1 nicht auf den wahren Wert eines festen Individuums auswirken. Die Wahrscheinlichkeitsverteilungen der Rater können unterschiedlich sein und damit auch die wahren Werte t_i bzw. t'_i, welche das Individuum Nr. i erhalten würde.

Als Folgerung ergibt sich:

Aus der linearen experimentellen Unabhängigkeit zweier ordinaler Tests folgt für ein festes Individuum die lokale Unkorreliertheit der Tests.

$$\underline{B.}\ \mathrm{Cov}_{aa'}[R_i, R'_i] = E_{aa'}[R_i R'_i] - t_i t'_i = E_{a'}[R'_i E_a[R_i | R'_i]] - t_i t'_i =$$

$$= E_{a'}[R'_i] E_a[R_i] - t_i t'_i$$

(wegen der linearen experimentellen Unabhängigkeit)

$$= t'_i t_i - t_i t'_i = 0.$$

1.2.3. Experimentelle Unabhängigkeit von ordinalen Tests

Eine stärkere Forderung an die Urteile zweier Rater als die der lokalen Unkorreliertheit und der linearen experimentellen Unabhängigkeit ist die experimentelle Unabhängigkeit der Raterurteile. Zwei ordinale Tests O und O', also zwei Rater, heißen *experimentell unabhängig*, wenn gilt:

$$P(\{R_1,\ldots,R_N\} = \{j_1,\ldots,j_N\}) \cap (\{R'_1,\ldots,R'_N\} = \{j'_1,\ldots,j'_N\})) =$$

$$= P(\{R_1,\ldots,R_N\} = \{j_1,\ldots,j_N\}) P(\{R'_1,\ldots,R'_N\} = \{j'_1,\ldots,j'_N\})$$

für beliebige Rangreihen der Rater.
Die Rater urteilen stochastisch unabhängig voneinander, beeinflussen sich also nicht beim Urteilen.

1.3. Gütemaße bei ordinalen Tests

1.3.1. Reliabilität von ordinalen Tests

Die *Reliabilität* eines Tests wird als Quadrat der Korrelation zwischen R und T definiert. Für die Korrelation erhält man:

$$\rho[R,T] = \frac{Cov_g[R,T]}{\sqrt{Var_g[R]Var_g[T]}} = \frac{E_g[RT]-E_g[R]E_g[T]}{\sqrt{Var_g[R]Var_g[T]}}$$

$$= \frac{E_g[(T+E)T]-(E_g[T])^2}{\sqrt{Var_g[R]Var_g[T]}} = \frac{E_g[T^2]-(E_g[T])^2}{\sqrt{Var_g[R]Var_g[T]}} = \frac{\sqrt{Var_g[T]}}{\sqrt{Var_g[R]}} .$$

Damit ist

$$\rho^2[R,T] = \frac{Var_g[T]}{Var_g[R]} = \frac{Var_g[T]}{(N^2-1)/12} .$$

Die Reliabilität ist bei ordinalen Tests nur von $Var_g[T]$ abhängig. Stellt ein Rater den ordinalen Test dar, so wird $\rho^2[R,T]$ sowohl von der Güte des Raterurteils als auch von den Individuen beeinflußt. Sind beispielsweise die Individuen nicht unterscheidbar, dann erhalten alle denselben wahren Wert und es ist $Var_g[T] = 0$. Damit ist $\rho^2[R,T] = 0$. Dieses ist auch der Fall, wenn die Individuen zwar unterscheidbar sind, der Rater aber nach einer Gleichverteilung urteilt. Jedes Individuum erhält dann mit der Wahrscheinlichkeit $\frac{1}{N}$ eine der N Rangzahlen, und es ist für i=1,...,N

$$E_a[R_i] = \sum_{j_1,\dots,j_N} j_i \cdot \frac{1}{N} = \frac{N+1}{2} .$$

In einem solchen Fall ist $Var_g[R] = Var_g[E]$.

Urteilt der Rater nach einer Einpunktverteilung, d.h., ist er sich seines Urteils völlig sicher, so sind die wahren Werte gleich den Rangzahlen und es ist $Var_g[R] = Var_g[T]$. Dann ist $\rho^2[R,T] = 1$.

1.3.2. Kriteriumsvalidität von Tests

Ein Rater beurteile die Individuen bezüglich eines Merkmals, ein weiterer Rater bezüglich einer Kriteriumsvariablen. Ein zufällig ausgewähltes Individuum bekomme bezüglich seiner Ausprägung auf der Kriteriumsvariablen die Rangzahl Q zugewiesen.

Die *Kriteriumsvalidität* zwischen einem ordinalen Test und einer lokal unkorrelierten Kriteriumsvariablen wird definiert als die Korrelation zwischen dem ordinalen Test und der Kriteriumsvariablen:

$$\rho[R,Q] = \frac{Cov_g[R,Q]}{\sqrt{Var_g[R] \cdot Var_g[Q]}} = \frac{Cov_g[R,Q]}{(N^2-1)/12} = \frac{Cov_g[T_R,T_Q]}{(N^2-1)/12} .$$

Ähnlich wie die Reliabilität, hängt auch die Validität von der Verteilung der wahren Werte ab. Die Validität ist Null, wenn $Cov_g[T_R,T_Q]=0$ ist. Das ist z.B. der Fall, wenn jeder der unabhängig urteilenden Rater seine Rangzahlen per Zufall zuordnet. Die Validität ist gleich 1, wenn $Cov_g[R,Q]=(N^2-1)/12$ ist. Dieses ist der Fall, wenn beide Rater mit Wahrscheinlichkeit 1 dieselben Rangreihen erstellen.

1.4. Äquivalente Messungen

Repräsentieren Rater die ordinalen Tests, so können die Anforderungen an den Urteilsprozeß unterschiedlich sein. Urteilen die Rater nach derselben Wahrscheinlichkeitsverteilung, so sollen die Tests als *Replikationen* bezeichnet werden. Sind nur Erwartungswerte und Varianzen der Verteilungen gleich, so wurden die Rater als homogen bezeichnet. Unabhängige homogene Rater sind z.B. zum Schätzen des wahren Wertes und der Varianz der Rangzahlen des Individuums i erforderlich. Sieht man vom Schätzen ab, so sind analog zur KTT auch hier für die Ableitungen schwächere Annahmen ausreichend.

Zwei lokal unkorrelierte Tests sollen *τ-äquivalent* heißen, wenn sie einem Individuum i, i=1,...,N, gleiche wahre Werte zuordnen:

$$E_a[R_i] = E_{a'}[R'_i] , \quad i=1,...,N .$$

Die Definition der KTT für wesentlich τ-äquivalente Tests kann nicht übernommen werden. Zwei Tests heißen *wesentlich τ-äquivalent*, wenn gilt

$$E_a[R_i] = E_{a'}[R'_i] + a , \quad i=1,...,N.$$

Dann müßte auch $E_g[R] = E_g[R'] + a$ sein, also $\frac{N+1}{2} = \frac{N+1}{2} + a$. Dieses ist nur für a=0 der Fall.

Schulman und Haden (1975) definieren allgemeiner T-ähnliche Tests:

Gegeben seien zwei lokal unkorrelierte Tests mit den Zufallsvariablen

T und T' für die wahren Werte. Zwei solche Tests heißen *T-ähnlich* (oder T'-ähnlich), wenn die wahren Werte des einen Tests sich als Lineartransformation aus den wahren Werten des anderen Tests berechnen lassen, wenn also $T' = aT + b$ $(T = cT'+d)$, $a > 0$ (bzw. $c > 0$) ist. Erhält also z.B. das Individuum Nr. i in dem einen Test den wahren Wert t_i, so erhält es in dem T-ähnlichen Test den wahren Wert (at_i+b). Es gilt

$$E_g[T'] = a\ E_g[T] + b\ ,$$

wobei

$$E_g[T'] = E_g[T] = \frac{N+1}{2}$$

ist.

Damit muß gelten

$$\frac{N+1}{2} = a\ \frac{N+1}{2} + b\ ,$$

$$b = \frac{N+1}{2}\ (1-a)\ .$$

a und b können also nicht unabhängig voneinander gewählt werden. Für die Varianzen gilt

$$Var_g[T'] = a^2 Var_g[T]\ .$$

Hat man einen Test und dazu einen T-ähnlichen Test zur Auswahl, so wird man den mit der höheren Reliabilität verwenden. Für die Reliabilität des T-ähnlichen Tests gilt:

$$\rho^2[R',T'] = \frac{Var_g[T']}{Var_g[R']} = \frac{a^2 Var_g[T]}{Var_g[R]} = a^2 \rho^2[R,T]\ .$$

Damit ist

$$a^2 = \frac{\rho^2[R',T']}{\rho^2[R,T]} = \frac{Var_g[T']}{Var_g[T]}\ .$$

Sind die Reliabilitäten der beiden Tests gleich, so ist $a^2 = 1$.

Für $\rho^2[R',T'] > \rho^2[R,T]$ ist $a^2 > 1$.

$a^2 > 1$ bedeutet, daß die wahren Werte beim T-ähnlichen Test eine größere Streuung aufweisen. In diesem Fall wird man diesen Test verwenden.

Ist $\rho^2[R',T'] < \rho^2[R,T]$, so ist $a^2 < 1$.

In diesem Falle wird man den ursprünglichen Test verwenden.

1.5. Paralleltests

In der KTT werden zwei lokal unkorrelierte Tests als parallel bezeichnet, wenn für ein festes Individuum i die wahren Werte und die intraindividuellen Varianzen gleich sind.

In Abweichung hierzu definieren Schulman und Haden (1975) *Paralleltests* als zwei lokal unkorrelierte Tests O, O' mit gleichen wahren Werten für ein Individuum i, i=1,...,N, d.h.

$$E_a[R_i]=E_a[R'_i],\ i=1,...,N\ .$$

Diese Forderung ist schwächer als diejenige der KTT. Es ist jetzt zwar $Var_g[E]=Var_g[E']$ (da $0=Var_g[R]-Var_g[R']=Var_g[T]-Var_g[T']+Var_g[E]-Var_g[E']=Var_g[E]-Var_g[E']$), jedoch folgt daraus nicht, daß für alle i $Var_a[E_i]=Var_a[E'_i]$ gilt.

Da die Gleichheit der Varianzen für ein festes Individuum nicht benötigt wird, soll auch hier die Definition Schulmans beibehalten werden. Diese Definition ist identisch mit derjenigen der τ-Äquivalenz von Tests in Abschnitt 1.4.

Paralleltests und Reliabilität

Da zwei Paralleltests lokal unkorreliert sind, weiterhin zu gleichen wahren Werten führen und Var[R] konstant ist, erhält man für die Korrelation zwischen zwei Paralleltests

$$\rho[R,R'] = \frac{Cov_g[R,R']}{\sqrt{Var_g[R]Var_g[R']}} = \frac{Cov_g[T,T']}{Var_g[R]} = \frac{Var_g[T]}{Var_g[R]} = \rho^2[R,T]\ .$$

Damit kann die Reliabilität als Korrelation zwischen den beiden Paralleltests berechnet werden, d.h. als Korrelation der Rangreihen zweier Rater. Die Reliabilität spiegelt hier gleichzeitig die Übereinstimmung der beiden Urteiler wider.

Damit ist

$$Var_g[T] = Var_g[R]\rho[R,R'] = \frac{N^2-1}{12}\ \rho[R,R']\ .$$

Mit

$$Var_g[T] + Var_g[E] = Var_g[R]$$

erhält man

$$Var_g[R] = Var_g[R]\rho[R,R'] + Var_g[E] ,$$

und

$$Var_g[E] = Var_g[R](1-\rho[R,R']) = \frac{N^2-1}{12}(1-\rho[R,R']) .$$

Ist $\rho[R,R']$ bekannt, so erhält man sowohl die Fehlervarianz als auch $Var_g[T]$.

m Paralleltests und ein lokal unkorreliertes Außenkriterium

Es seien m Paralleltests und ein lokal unkorreliertes Außenkriterium gegeben. Für ein zufällig ausgewähltes Individuum seien die Rangzahlen bzw. die wahren Werte aus den Paralleltests bezeichnet mit $R',R'',\ldots,R^{(m)}$ bzw. $T',T'',\ldots,T^{(m)}$. Rangzahl bzw. wahrer Wert bezüglich des Außenkriteriums seien Q bzw. T_Q.
Dann gilt

1. Alle Korrelationen der parallelen Tests sind gleich:
$$\rho[R',R''] = \rho[R',R'''] = \ldots = \frac{Var_g[T]}{Var_g[R]} .$$

2. Alle parallelen Tests korrelieren gleich mit dem Außenkriterium:
$$\rho[R',Q] = \rho[R'',Q] = \ldots = \rho[R^{(m)},Q] .$$

$\underline{B}_1$. Da für die Paralleltests $T'=T''=\ldots=T^{(m)}=T$ gilt, und die Tests lokal unkorreliert sind, gilt

$Cov_g[R',R''] = Cov_g[T',T''] = Var_g[T],\ldots$ usw.

Mit $Var_g[R'] = Var_g[R''] = \ldots = (N^2-1)/12$ sind alle Korrelationen gleich.

$\underline{B}_2$. Da die Paralleltests und das Außenkriterium als lokal unkorreliert vorausgesetzt wurden, ist

$Cov_g[R',Q] = Cov_g[T',T_Q],\ldots$ usw.

Da weiterhin $T'=T''=\ldots=T^{(m)}=T$ gilt, ist

$$Cov_g[T',T_Q] = Cov_g[T'',T_Q] = \ldots = Cov_g[T,T_Q] .$$

Mit

$$\rho[R',Q] = \frac{Cov_g[R',Q]}{\sqrt{Var_g[R]Var_g[Q]}} = \frac{Cov_g[T,T_Q]}{\sqrt{Var_g[R]Var_g[Q]}}$$

sind damit auch die Korrelationen gleich:

$$\rho[R',Q] = \rho[R'',Q] = \ldots = \rho[R^{(m)},Q] .$$

1.6. Attenuationsformeln

1. Es seien zwei lokal unkorrelierte Tests gegeben. Dann gilt

$$\rho[T,T'] = \frac{\rho[R,R']}{\sqrt{\rho^2[R,T]\cdot\rho^2[R',T']}} \quad , \quad \rho^2[R,T] > 0, \; \rho^2[R',T'] > 0 .$$

Die Formel gibt die Beziehung zwischen der Korrelation der wahren Werte der beiden Tests in Abhängigkeit von der Korrelation zwischen den Rangzahlen und den Reliabilitäten an.

B. $$\frac{\rho[R,R']}{\sqrt{\rho^2[R,T]\cdot\rho^2[R',T']}} = \frac{Cov_g[R,R']}{\sqrt{Var_g[R]\cdot Var_g[R']}} \cdot \frac{\sqrt{Var_g[R]\cdot Var_g[R']}}{\sqrt{Var_g[T]\cdot Var_g[T']}} =$$

$$= \frac{Cov_g[T,T']}{\sqrt{Var_g[T]\cdot Var_g[T']}} = \rho[T,T'] .$$

Löst man die oben angegebene Formel nach $\rho[R,R']$ auf, so ergibt sich

$$\rho[R,R'] = \rho[T,T']\cdot\sqrt{\rho^2[R,T]\rho^2[R',T']} .$$

Für gegebenes $\rho[T,T']$ und $\rho^2[R',T']$ steigt bei wachsender Reliabilität $\rho^2[R,T]$ auch $\rho[R,R']$ monoton an. Handelt es sich bei R' um die Rangzahl aus einem Außenkriterium, so stellt $\rho[R,R']$ die Validität dar. Die Formel besagt somit auch, daß die Validität monoton mit der Reliabilität steigt. Da $\rho[T,T']$ und $\rho^2[R',T']$ maximal gleich 1 werden können, kann die Validität nicht größer als $\sqrt{\rho^2[R,T]}$ werden:

$$|\rho[R,R']| = |\rho[T,T']|\cdot\sqrt{\rho^2[R,T]\rho^2[R',T']}$$

$$|\rho[R,R']| \leq \sqrt{\rho^2[R,T]} .$$

In Abschnitt 1.9.2. soll die Veränderung dieser Beziehung betrachtet werden, wenn die lokale Unkorreliertheit der Rater nicht gegeben ist.

2. Es seien zwei lokal unkorrelierte Tests gegeben (Rangzahlen R bzw. Q) und zu jedem der Tests ein Paralleltest (Rangzahlen R' bzw. Q'). Dann gilt:

$$\rho[R,T_Q] = \frac{\rho[R,Q]}{\sqrt{\rho[Q,Q']}} \quad , \quad \rho[Q,Q'] \neq 0 \ ,$$

$$\rho[Q,T_R] = \frac{\rho[R,Q]}{\sqrt{\rho[R,R']}} \quad , \quad \rho[R,R'] \neq 0 \ .$$

Diese Formeln geben die Korrelation zwischen der Rangzahl in einem Test und dem wahren Wert in einem zweiten Test an in Abhängigkeit von der Korrelation der Rangzahlen beider Tests und der Reliabilität des zweiten Tests.

B. $$\frac{\rho[R,Q]}{\sqrt{\rho[Q,Q']}} = \frac{Cov_g[R,Q]}{\sqrt{Var_g[R] \cdot Var_g[Q]}} \ \frac{1}{\sqrt{\rho^2[Q,T_Q]}} = \frac{Cov_g[T_R,T_Q]}{\sqrt{Var_g[R] \cdot Var_g[T_Q]}} =$$

$$= \rho[R,T_Q] \ .$$

Dabei wurde $Cov_g[R,T_Q] = Cov_g[T_R,T_Q]$ benutzt.

Die Formeln 1. und 2. werden als *Attenuationsformeln* bezeichnet.

1.7. Zusammensetzen von Tests

Es seien k ordinale Tests also z.B. k Rater gegeben. Jeder der Rater beurteilt die Individuen. Für ein zufällig ausgewähltes Individuum mögen sich die Rangzahlen $R^1, R^2, \ldots, R^k$ ergeben. Durch Addition wird ein Gesamtscore X gebildet:

$$X = \sum_{j=1}^{k} R^j \ .$$

Dann gilt für den wahren Wert und den Meßfehler des so zusammengesetzten Tests

$$T_X = \sum_{j=1}^{k} T^j \ , \quad E_X = \sum_{j=1}^{k} E^j \ .$$

1.7.1. Zusammensetzen von Paralleltests und Veränderung der Reliabilitä

Es beurteilen z.B. k parallele Rater die N Individuen bezüglich desselben Merkmals. Um zu zeigen, daß die Formel von Spearman-Brown auch für zusammengesetzte Tests gilt, werden zunächst die Varianzen für X, T_X und E_X, sowie $Cov_g[X,T_X]$ berechnet:

$$Var_g[X] = k\ Var_g[R](1+(k-1)\rho[R,R']) = k\ \frac{N^2-1}{12}\ (1+(k-1)\rho[R,R'])\ ,$$

$Var_g[T_X] = Var_g[kT] = k^2 Var_g[T]$, mit $T^1=T^2=\ldots T^k=T$, da es sich um Parallelteste handelt.

$$Var_g[E_X] = Var_g[E^1] + Var_g[E^2] + \ldots + Var_g[E^k] =$$

$$= k\ (\frac{N^2-1}{12} - Var_g[T]) = k\ Var_g[E]\ .$$

$$Cov_g[X,T_X] = E_g[X\cdot T_X] - E_g[X]\cdot E_g[T_X] = E_g[(T_X+E_X)T_X] - E_g[T_X+E_X]E_g[T_X]$$

$$= E_g[T_X^2] - (E_g[T_X])^2 = Var_g[T_X] = k^2 Var_g[T]\ .$$

Für die Reliabilität dieses zusammengesetzten Tests gilt

$$\rho^2[X,T_X] = \frac{k\rho^2[R,T]}{1+(k-1)\rho^2[R,T]}\ .$$

B. $$\rho^2[X,T_X] = \frac{(Cov_g[X,T_X])^2}{Var_g[X]\cdot Var_g[T_X]} = \frac{(k^2 Var_g[T])^2}{kVar_g[R](1+(k-1)\rho[R,R'])\cdot k^2 Var_g[T]} =$$

$$= \frac{k^2 Var_g[T]/Var_g[R]}{k(1+(k-1)\rho[R,R'])} =$$

$$= \frac{k\ \rho^2[R,T]}{1+(k-1)\rho^2[R,T]}\ .$$

Damit gilt auch hier die aus der KTT bekannte allgemeine *Formel von Spearman-Brown* für die Reliabilität eines Tests, welcher aus k parallelen Komponenten zusammengesetzt wurde. Man kann hieraus ersehen, daß die Reliabilität des zusammengesetzten Tests größer ist als die des einfachen. Die Meßgenauigkeit wächst also, je mehr parallele Rater die Individuen beurteilen.

Für k=2 parallele Komponenten erhält man

$$\rho^2[X,T_X] = \frac{2\rho^2[R,T]}{1+\rho^2[R,T]}\ .$$

1.7.2. Zusammensetzen lokal unkorrelierter Tests und Veränderung der Reliabilität

Es werde nur vorausgesetzt, daß die Rater, welche die ordinalen Tests repräsentieren, lokal unkorreliert voneinander urteilen. So kann es sich z.B. bei dem Zusammensetzen der Tests jetzt auch um das Bilden

eines allgemeinen Scores handeln, wobei die Rater die Individuen bezüglich unterschiedlicher Merkmale beurteilen. Die Rater sind nicht mehr parallel.

In der KTT wird eine untere Schranke für die Reliabilität eines solchen zusammengesetzten Tests angegeben.

Diese lautet, übertragen für k=2 lokal unkorrelierte ordinale Tests:

$$\rho^2[X,T_X] \geq 2\left[1 - \frac{Var_g[R]+Var_g[R']}{Var_g[X]}\right] .$$

Die Ungleichung kann vereinfacht werden zu

$$\rho^2[X,T_X] \geq \frac{4Cov_g[T,T']}{Var_g[X]} ,$$

$$\frac{Var_g[T_X]}{Var_g[X]} \geq \frac{4Cov_g[T,T']}{Var_g[X]} ,$$

$$Var_g[T_X] \geq 4Cov_g[T,T'] .$$

Diese zuletzt angegebene Ungleichung wird bewiesen.

B. Es gilt

$$Var_g[T-T'] = Var_g[T]+Var_g[T']-2Cov_g[T,T'] \geq 0 .$$

Damit ist $Var_g[T]+Var_g[T'] \geq 2Cov_g[T,T']$.

Man erhält dann

$$Var_g[T_X] = Var_g[T+T'] = Var_g[T]+Var_g[T']+2Cov_g[T,T'] \geq 4Cov_g[T,T'].$$

Für zwei parallele Komponenten erhält man mit
$Var_g[T_X] = 4Var_g[T] = 4Cov_g[T,T']$

$$\rho^2[X,T_X] = \frac{4Cov_g[T,T']}{Var_g[X]} = \frac{4Var_g[T]}{2Var_g[R](1+\rho^2[R,T])} = \frac{2Var_g[T]/Var_g[R]}{1+\rho^2[R,T]} =$$

$$= \frac{2\rho^2[R,T]}{1+\rho^2[R,T]} .$$

Dieses ist die Formel von Spearman-Brown (1.7.1.), welche sich hier als Spezialfall ergibt.

Sollen k lokal unkorrelierte Tests zusammengesetzt werden, so lautet die untere Schranke für die Reliabilität des zusammengesetzten ordinalen Tests

$$\rho^2[X,T_X] \geq \frac{k}{k-1}\left[1 - \frac{\sum_{j=1}^{k} Var_g[R^j]}{Var_g[X]}\right],$$

wobei die rechte Seite als Koeffizient α (Cronbach, 1951) bekannt ist. Wegen $Var_g[R^j] = Var_g[R] = (N^2-1)/12$, $j=1,\ldots,k$, hat α die Form

$$\alpha = \frac{k}{k-1}\left[1 - \frac{kVar_g[R]}{kVar_g[R]+2\sum_{j<l} Cov_g[R^j,R^l]}\right] = \frac{k}{k-1}\,\frac{2\sum_{j<l} Cov_g[R^j,R^l]}{Var_g[X]}.$$

Mit

$$\rho^2[X,T_X] = \frac{Var_g[T_X]}{Var_g[X]}$$

kann die Ungleichung vereinfacht werden und die zu beweisende Behauptung lautet:

$$Var_g[T_X] \geq \frac{k}{k-1} \cdot 2\sum_{j<l} Cov_g[R^j,R^l].$$

B. Unter Verwendung von

$$Var_g[T_X] = Var_g[T^1] + \ldots + Var_g[T^k] + 2\sum_{j<l} Cov_g[T^j,T^l]$$

und

$$Var_g[T^j] + Var_g[T^l] \geq 2Cov_g[T^j,T^l]$$

ist

$$Var_g[T_X] = \frac{1}{k-1}\sum_{j=1}^{k}\sum_{\substack{l=1\\ j<l}}^{k}(Var_g[T^j] + Var_g[T^l]) + 2\sum_{j<l} Cov_g[T^j,T^l] \geq$$

$$\geq \frac{2}{k-1}\sum_{j<l} Cov_g[T^j,T^l] + 2\sum_{j<l} Cov_g[T^j,T^l] =$$

$$= \sum_{j<l} Cov_g[T^j,T^l]\left\{\frac{2}{k-1} + 2\right\} = \frac{2k}{k-1}\sum_{j<l} Cov_g[T^j,T^l].$$

Damit wurde die Ungleichung bewiesen.

Für k parallele Komponenten erhält man als Spezialfall statt der Ungleichung eine Gleichung:

$$\rho^2[X,T_X] = \frac{k}{k-1}\,\frac{2\binom{k}{2}\,\mathrm{Var}_g[T]}{k\mathrm{Var}_g[R]\,(1+(k-1)\rho^2[R,T])} = \frac{k\mathrm{Var}_g[T]/\mathrm{Var}_g[R]}{1+(k-1)\rho^2[R,T]} =$$

$$= \frac{k\rho^2[R,T]}{1+(k-1)\rho^2[R,T]} .$$

Dieses ist die Spearman-Brown-Formel für k parallele Komponenten.

1.8. Veränderung der Validität beim Zusammensetzen von Tests

Es werden k parallele Tests zusammengesetzt: $X = R^1+R^2+\ldots+R^k$. Ein zu X lokal unkorreliertes Außenkriterium sei aus l parallelen Komponenten zusammengesetzt: $Y = Q^1+\ldots+Q^l$. Nach 1.5. korrelieren alle parallelen Tests $R^1,\ldots,R^k$ gleich mit jedem der $Q^1,\ldots,Q^l$:

$$\rho[R^1,Q^1] = \rho[R^1,Q^2] = \ldots =: \rho[R,Q] .$$

Die Reliabilität eines Einzeltests von X werde mit $\rho[R,R']$ bezeichnet, die Reliabilität eines Einzeltests von Y mit $\rho[Q,Q']$.

Es gilt für die Validität:

$$\rho[X,Y] = \frac{\sqrt{k\cdot l}\cdot\rho[R,Q]}{\sqrt{(1+(k-1)\rho[R,R'])\,(1+(l-1)\rho[Q,Q'])}} .$$

B. $$\rho[X,Y] = \frac{\mathrm{Cov}_g[X,Y]}{\sqrt{\mathrm{Var}_g[X]}\sqrt{\mathrm{Var}_g[Y]}}$$

$$= \frac{\mathrm{Cov}_g[R^1+\ldots+R^k,Q^1+\ldots+Q^l]}{\sqrt{k\cdot\mathrm{Var}_g[R]\,(1+(k-1)\rho[R,R'])}\cdot\sqrt{l\mathrm{Var}_g[Q]\,(1+(l-1)\rho[Q,Q'])}}$$

$$= \frac{l\cdot k\cdot\mathrm{Cov}_g[R,Q]}{\sqrt{k\cdot l\cdot\mathrm{Var}_g[R]\cdot\mathrm{Var}_g[Q]}\cdot\sqrt{(1+(k-1)\rho[R,R'])}\cdot\sqrt{(1+(l-1)\rho[Q,Q'])}}$$

$$= \frac{\sqrt{l\cdot k}\,\rho[R,Q]}{\sqrt{1+(k-1)\rho[R,R']}\cdot\sqrt{1+(l-1)\rho[Q,Q']}} .$$

Aus der Formel kann man ersehen, wie die Validität durch Zusammensetzen der Tests wächst. Für l=1 (1 Rater beurteilt die Individuen

bezüglich des Außenkriteriums) erhält man

$$\rho[X,Y] = \frac{\sqrt{k}\rho[R,Q]}{\sqrt{1+(k-1)\rho[R,R']}} .$$

Sind X und Y Paralleltests, so kann die Reliabilität, berechnet aus der Korrelation dieser Paralleltests, berechnet werden. Für $1=k$ erhält man

$$\rho[X,Y] = \frac{k\rho[R,R']}{1+(k-1)\rho[R,R']} .$$

Hieraus sieht man - analog zur KTT -, daß die Validität mit der Testlänge k langsamer ansteigt als die Reliabilität.

($\frac{\sqrt{k}}{\sqrt{1+(k-1)\rho[R,R']}}$ steigt langsamer als $\frac{k}{1+(k-1)\rho[R,R']}$).

Man erhält für die zusammengesetzten Tests X,Y die *Attenuationsformeln*:

1. $\rho[T_X,T_Y] = \frac{\rho[R,Q]}{\sqrt{\rho[R,R']}\sqrt{\rho[Q,Q']}}$, $\rho[R,R'] \neq 0$, $\rho[Q,Q'] \neq 0$.

Die Formel gibt die Beziehung zwischen der Korrelation der wahren Werte der beiden zusammengesetzten Tests in Abhängigkeit von der Korrelation zwischen den Rangzahlen zweier einzelner Tests und den Reliabilitäten dieser Tests an.

B. Es gilt (siehe 1.6.)

$$\rho[T_X,T_Y] = \frac{\rho[X,Y]}{\sqrt{\rho[X,X']\rho[Y,Y']}}$$

$$= \frac{\sqrt{1\cdot k}\rho[R,Q]\cdot\sqrt{1+(k-1)\rho[R,R']}\sqrt{1+(1-1)\rho[Q,Q']}}{\sqrt{1+(k-1)\rho[R,R']}\cdot\sqrt{1+(1-1)\rho[Q,Q']}\sqrt{k\rho[R,R']}\cdot\sqrt{1\rho[Q,Q']}}$$

$$= \frac{\sqrt{1\cdot k}\rho[R,Q]}{\sqrt{k\rho[R,R']}\sqrt{1\rho[Q,Q']}}$$

$$= \frac{\rho[R,Q]}{\sqrt{\rho[R,R']\rho[Q,Q']}} .$$

2. $\rho[R,T_Y] = \frac{\rho[R,Q]}{\sqrt{\rho[Q,Q']}}$, $\rho[Q,Q'] \neq 0$,

$\rho[T_X,Q] = \frac{\rho[R,Q]}{\sqrt{\rho[R,R']}}$, $\rho[R,R'] \neq 0$.

B. Es gilt (siehe 1.6.)

$$\rho[R,T_Y] = \frac{\rho[R,Y]}{\sqrt{\rho[Y,Y']}} = \frac{\sqrt{1}\rho[R,Q]\cdot\sqrt{1+(1-1)\rho[Q,Q']}}{\sqrt{1+(1-1)\rho[Q,Q']}\cdot\sqrt{1\rho[Q,Q']}}$$

$$= \frac{\sqrt{1}\rho[R,Q]}{\sqrt{1\rho[Q,Q']}} = \frac{\rho[R,Q]}{\sqrt{\rho[Q,Q']}} .$$

Die andere Formel wird analog bewiesen.

Für die beiden Formeln werden zwei Tests betrachtet, von denen der eine aus nur 1 Komponente, der andere aus 1 bzw. k parallelen Komponenten besteht. Die Formeln geben die Beziehung an zwischen der Korrelation der Rangzahlen in einem Test und den wahren Werten eines anderen, zusammengesetzten Tests in Abhängigkeit von der Korrelation der Rangzahlen in beiden Tests und der Reliabilität des zusammengesetzten Tests.
Die Attenuationsformeln 1. und 2. sind von derselben Form wie bei nichtzusammengesetzten Tests.

1.9. Lokal korrelierte ordinale Tests

Wenn bisher von mehreren ordinalen Tests gesprochen wurde, so handelte es sich stets um mehrere Rater, welche als lokal unkorreliert bzw. sogar als unabhängig voneinander urteilend vorausgesetzt wurden. So wurden 2 Paralleltests als 2 lokal unkorreliert urteilende Rater definiert, deren Urteile für ein festes Individuum i, i=1,...,N, zu gleichen wahren Werten führen. Auch zur Bestimmung der Validität wurden 2 Rater benötigt, von denen der eine die Individuen bezüglich des Merkmals, der andere dieselben Individuen bezüglich des Kriteriums beurteilten.

Beurteilt derselbe Rater die Individuen bezüglich des Merkmals und bezüglich des Kriteriums, so können die entstehenden Rangreihen korrelierte Fehler aufweisen. Eine Korrelation zwischen den Fehlern kann auch vorhanden sein, wenn zwei Rater urteilen, die sich gegenseitig beeinflussen oder wenn ein Rater zweimal dieselben Indivi-

duen beurteilt. Wendet man bei Verletzung der lokalen Unkorreliertheit die Formeln an, welche für lokal unkorrelierte Tests gelten, so können beträchtliche Fehler auftreten.

Für die KTT betrachteten Zimmerman und Williams (1977) und Williams und Zimmerman (1982) u.a. Validität, Attenuation und Testverlängerung bei Tests mit korrelierten Fehlern. Diese Ergebnisse sind auch gültig, wenn sie auf Rangzahlen übertragen werden. Repräsentieren Rater die ordinalen Tests, so ist es im Unterschied zur KTT leichter, korrelierte Fehler zu vermeiden. Aus Mangel an Ratern oder aus Unkenntnis der möglichen Fehler geschieht es jedoch, daß in praktischen Situationen ein Rater Individuen bezüglich mehrerer Merkmale beurteilt. Wie die folgenden Ausführungen zeigen, ist von diesem Vorgehen abzuraten.

1.9.1. Reliabilität

Es werde angenommen, daß die Reliabilität eines ordinalen Tests als Korrelation zweier ordinaler Tests, welche für die Individuen gleiche wahre Werte liefern, aber korrelierte Fehler aufweisen, bestimmt werden soll. Durch die Annahme einer Korrelation zwischen den Fehlern handelt es sich nicht mehr um parallele Tests im oben definierten Sinn. Angenommen, ein Rater repräsentiert den ordinalen Test, dessen Reliabilität bestimmt werden soll. Beurteilt dieser Rater zweimal die Individuen mit dem Ziel, die Reliabilität zu bestimmen, so kann die lokale Unkorreliertheit verletzt sein. Man erhält dann:

$$\rho[R,R'] = \frac{Cov_g[R,R']}{Var_g[R]} = \frac{Cov_g[T,T']+Cov_g[E,E']}{Var_g[R]} =$$

$$= \frac{Var_g[T]}{Var_g[R]} + \frac{Cov_g[E,E']}{Var_g[R]} = \rho^2[R,T] + \rho[E,E'] \frac{Var_g[E]}{Var_g[R]} .$$

Für $\rho[E,E'] = 0$ erhält man $\rho[R,R'] = \rho^2[R,T]$, d.h. man hat die Reliabilität als Korrelation zwischen lokal unkorrelierten Paralleltests (parallelen Ratern).
Je nachdem, ob $\rho[E,E']$ positiv oder negativ ist, ist $\rho[R,R']$ größer als $\rho^2[R,T]$ oder kleiner. Der Einfluß von $\rho[E,E']$ auf $\rho[R,R']$ ist gering, wenn $Var_g[E]$ nahe Null ist. Das ist der Fall, wenn die Reliabilität $\rho^2[R,T]$ des Raters hoch ist. Ist jedoch die Reliabilität des Raters nicht sehr hoch, dann ist $Var_g[T]$ klein, d.h. $Var_g[E]$

hoch, und $\rho[R,R']$ kann beträchtlich von $\rho^2[R,T]$ abweichen.
Die Bestimmung der Reliabilität eines Tests (Raters) als Korrelation zwischen Paralleltests im klassischen Sinne, wobei jedoch derselbe Rater zweimal die Individuen beurteilt, ist also nur anzuraten, wenn die Reliabilität des Raters hoch ist. Dann werden sich $\rho[R,R']$ und $\rho^2[R,T]$ nur geringfügig voneinander unterscheiden. D.h. jedoch, daß über die Reliabilität schon etwas bekannt sein muß, was in der praktischen Anwendung kaum der Fall sein wird. Damit ist dieses Vorgehen nicht zu empfehlen.

1.9.2. Attenuation

Gegeben sei ein ordinaler Test (Rangzahl R) und ein Außenkriterium (Rangzahl Q).
Beurteilt derselbe Rater die N Individuen sowohl bezüglich des Merkmals als auch bezüglich des Kriteriums, so ist möglicherweise die lokale Unkorreliertheit verletzt. Dann ist

$$\mathrm{Cov}_g[R,Q] = \mathrm{Cov}_g[T_R,T_Q] + \mathrm{Cov}_g[E_R,E_Q] .$$

Zur Bestimmung der Validität wird dividiert durch $\sqrt{\mathrm{Var}_g[R]\mathrm{Var}_g[Q]}$:

$$\frac{\mathrm{Cov}_g[R,Q]}{\sqrt{\mathrm{Var}_g[R]\mathrm{Var}_g[Q]}} = \frac{\mathrm{Cov}_g[T_R,T_Q]}{\sqrt{\mathrm{Var}_g[R]\mathrm{Var}_g[Q]}} + \frac{\mathrm{Cov}_g[E_R,E_Q]}{\sqrt{\mathrm{Var}_g[R]\mathrm{Var}_g[Q]}}$$

$$\rho[R,Q] = \rho[T_R,T_Q]\frac{\sqrt{\mathrm{Var}_g[T_R]\mathrm{Var}_g[T_Q]}}{\sqrt{\mathrm{Var}_g[R]\mathrm{Var}_g[Q]}} + \rho[E_R,E_Q]\frac{\sqrt{\mathrm{Var}_g[E_R]\mathrm{Var}_g[E_Q]}}{\sqrt{\mathrm{Var}_g[R]\mathrm{Var}_g[Q]}}$$

$$\rho[R,Q] = \rho[T_R,T_Q]\cdot\sqrt{\rho^2[R,T_R]\rho^2[Q,T_Q]} + \rho[E_R,E_Q]\sqrt{(1-\rho^2[R,T_R])(1-\rho^2[Q,T_Q])} .$$

Löst man diese Gleichung nach $\rho[T_R,T_Q]$ auf, so erhält man

$$\rho[T_R,T_Q] = \frac{\rho[R,Q]}{\sqrt{\rho^2[R,T_R]\rho^2[Q,T_Q]}} - \rho[E_R,E_Q]\frac{\sqrt{(1-\rho^2[R,T_R])(1-\rho^2[Q,T_Q])}}{\sqrt{\rho^2[R,T_R]\rho^2[Q,T_Q]}} ,$$

$$\rho^2[R,T_R] > 0 \ , \ \rho^2[Q,T_Q] > 0 .$$

Für $\rho[E_R,E_Q] = 0$ ergibt sich die bekannte Attenuationsformel für lokal unkorrelierte Tests (siehe 1.6.):

$$\rho[T_R,T_Q] = \frac{\rho[R,Q]}{\sqrt{\rho^2[R,T_R]\rho^2[Q,T_Q]}} .$$

Wird der Einfachheit halber $\rho^2[R,T_R] = \rho^2[Q,T]$ angenommen, so ist

$$\rho[T_R,T_Q] = \frac{\rho[R,Q]}{\rho^2[R,T_R]} - \rho[E_R,E_Q]\,\frac{1-\rho^2[R,T_R]}{\rho^2[R,T_R]} \quad .$$

Ist z.B. $\rho^2[R,T_R]$ sehr klein und $\rho[E_R,E_Q] \neq 0$, so können beträchtliche Abweichungen im Vergleich zu lokal unkorrelierten Ratern entstehen. Ist die Reliabilität $\rho^2[R,T_R]$ hoch, so ist $1-\rho^2[R,T_R]$ nahe Null und $\rho[T_R,T_Q]$ wird bei positiv korrelierten Fehlern nur geringfügig kleiner sein als bei unkorrelierten Ratern.
Ist jedoch $\rho^2[R,T_R]$ hoch, so ist eigentlich die Attenuationsformel überflüssig, da sie eine Korrektur darstellt, falls die Reliabilitäten der Rater nicht sehr hoch sind.

1.9.3. Kriteriumsvalidität

Die eben abgeleitete Formel

$$\rho[R,Q] = \rho[T_R,T_Q]\sqrt{\rho^2[R,T_R]\rho^2[Q,T_Q]} + \rho[E_R,E_Q]\cdot\sqrt{(1-\rho^2[R,T_R])\,(1-\rho^2[Q,T_Q])}$$

stellt die Formel für die Validität lokal korrelierter ordinaler Tests dar.

Ist $\rho[E_R,E_Q] = 0$, so erhält man für die Validität lokal unkorrelierter Tests

$$\rho[R,Q] = \rho[T_R,T_Q]\sqrt{\rho^2[R,T_R]\;\rho^2[Q,T_Q]} \quad .$$

Für gegebenes $\rho[T_R,T_Q]$ und $\rho^2[Q,T_Q]$ steigt bei wachsender Reliabilität $\rho^2[R,T_R]$ die Validität $\rho[R,Q]$ monoton an. Weiterhin gilt

$$|\rho[R,Q]| \leq \sqrt{\rho^2[R,T_R]} \quad .$$

Dieses muß beim Vorliegen korrelierter Fehler nicht gelten.
Die Beziehung zwischen Reliabilität und Validität kann dann monoton steigend, monoton fallend oder nichtmonoton sein. Die Validität kann in diesem Fall sogar größer als die Wurzel aus der Reliabilität werden.

Um dieses zu zeigen, werde die Validität als Funktion der Reliabilität betrachtet:
$\rho[R,Q] = f(\rho^2[R,T_R])$.

$\rho^2[Q,T_Q]$, $\rho[T_R,T_Q]$ und $\rho[E_R,E_Q]$ seien Parameter. Um etwas über den

Verlauf der Funktion f im Intervall (0,1) aussagen zu können, werden die Extremwerte bestimmt. Hierzu wird zunächst die 1. Ableitung gebildet:

$$f' = \frac{d\rho[R,Q]}{d\rho^2[R,T_R]} = \rho[T_R,T_Q] \frac{\rho^2[Q,T_Q]}{2\sqrt{\rho^2[R,T_R]\rho^2[Q,T_Q]}} -$$

$$- \rho[E_R,E_Q] \frac{1-\rho^2[Q,T_Q]}{2\sqrt{(1-\rho^2[R,T_R])(1-\rho^2[Q,T_Q])}}, \qquad 0 < \rho^2[R,T_R] < 1 ,$$

$$0 < \rho^2[Q,T_Q] < 1 .$$

Durch Nullsetzen und Auflösen nach $\rho^2[R,T_R] = m$ erhält man den Wert

$$m = \frac{\rho^2[T_R,T_Q]\rho^2[Q,T_Q]}{\rho^2[T_R,T_Q]\rho^2[Q,T_Q]+\rho^2[E_R,E_Q](1-\rho^2[Q,T_Q])} ,$$

an welchem f einen Extremwert annimmt.

Die zweite Ableitung gibt an, um welche Art von Extremwert es sich handelt.

$$f'' = \frac{d^2\rho[R,Q]}{d(\rho^2[R,T_R])^2} = - \rho[T_R,T_Q] \frac{\sqrt{\rho^2[Q,T_Q]}}{4\sqrt{\rho^2[R,T_R]^3}} - \rho[E_R,E_Q] \frac{\sqrt{1-\rho^2[Q,T_Q]}}{4\sqrt{(1-\rho^2[R,T_R])^3}} .$$

Folgende Fälle können unterschieden werden:

1. $\rho[T_R,T_Q]>0$, $\rho[E_R,E_Q]>0$.

   Es ist $f''<0$, d.h. $f(\rho^2[R,T_R])$ hat an der Stelle m ein Maximum.

2. $\rho[T_R,T_Q]<0$, $\rho[E_R,E_Q]<0$.

   Es ist $f''>0$, d.h. $f(\rho^2[R,T_R])$ hat an der Stelle m ein Minimum.

3. $\rho[T_R,T_Q]>0$, $\rho[E_R,E_Q]<0$ oder umgekehrt.

   Nullsetzen der zweiten Ableitung und Auflösen nach $\rho^2[R,T_R]$ ergibt einen Wendepunkt von f in

$$w = \frac{\{\rho^2[T_R,T_Q]\rho^2[Q,T_Q]\}^{1/3}}{\{\rho^2[T_R,T_Q]\rho^2[Q,T_Q]\}^{1/3} + \{\rho^2[E_R,E_Q](1-\rho^2[Q,T_Q])\}^{1/3}} .$$

Im Fall 1 steigt die Validität, wenn die Reliabilität von 0 auf m

ansteigt, erreicht im Punkt m ihr Maximum und fällt wieder, wenn die Reliabilität von m auf 1 wächst. Umgekehrt verhält es sich im Fall 2. Hier fällt die Validität, wenn die Reliabilität auf m anwächst und steigt dann an, wenn die Reliabilität von m auf 1 ansteigt.

Fall 3 bedeutet, daß die Funktion f (Validität) ihr Krümmungsverhalten ändert, wenn die Reliabilität den Wert w annimmt. Die Validität als Funktion der Reliabilität ist hier im gesamten Intervall (0,1) monoton steigend, wenn $\rho[T_R,T_Q]>0$ und $\rho[E_R,E_Q]<0$ und monoton fallend für $\rho[T_R,T_Q]<0$ und $\rho[E_R,E_Q]>0$.

Mit diesen Betrachtungen wurde gezeigt, daß im Falle korrelierter Fehler die Validität nicht in jedem Falle monoton mit der Reliabilität wächst, sondern daß ein Maximum bzw. Minimum im Intervall (0,1) vorliegen kann. Die Lage des Extremwertes ist auch von $\rho[E_R,E_Q]$ abhängig. Für $\rho[E_R,E_Q]=0$ ist m=1. Für wachsendes $\rho[E_R,E_Q]$ wandert m auf der x-Achse (Reliabilität) nach links.

Daß die Validität einen größeren Wert als $\sqrt{\rho^2[R,T_R]}$ annehmen kann, zeigt die folgende Betrachtung. Ausgehend von

$$\rho[R,Q]=\rho[T_R,T_Q]\sqrt{\rho^2[R,T_R]\rho^2[Q,T_Q]}+\rho[E_R,E_Q]\sqrt{(1-\rho^2[R,T_R])(1-\rho^2[Q,T_Q])}$$

setzt man für $\rho[T_R,T_Q]$ und $\rho^2[Q,T_Q]$ die maximalen Werte 1 bzw. m' ein (wobei m' analog zu m gewonnen wird) und erhält für $\rho[R,Q]$:

$$\rho[R,Q]=\sqrt{\rho^2[R,T_R]+\rho^2[E_R,E_Q](1-\rho^2[R,T_R])}\ .$$

Dieser Ausdruck gibt den maximalen Wert an, welchen die Validität in Abhängigkeit von der Reliabilität annehmen kann. Für $\rho[E_R,E_Q]=0$ erhält man als klassisches Ergebnis $\rho[R,Q]=\sqrt{\rho^2[R,T_R]}$. Für $\rho^2[E_R,E_Q]>0$ ergibt sich $\rho[R,Q]>\sqrt{\rho^2[R,T_R]}$.

Die geringsten Abweichungen vom Fall lokal unkorrelierter Rater ergeben sich auch hier, wenn sowohl die Reliabilität des Tests wie die des Kriteriums hoch sind.

2. Schätzen von Parametern

Für die bisherigen Ableitungen wurde angenommen, daß N Individuen (Population) in eine Rangreihe gebracht werden und daß die Wahrscheinlichkeitsverteilung der möglichen Rangordnungen bekannt ist.

Die Verteilung der Rangordnungen wird in der Regel nicht bekannt sein. Weiterhin ist es möglich, daß statt der N Individuen der Population nur eine Zufallsstichprobe von n Individuen aus der Population in eine Rangreihe gebracht werden sollen. Beispielsweise ist die Population zu groß und es ist nicht möglich, alle Individuen in eine Rangreihe zu bringen. Im folgenden werden 2 Situationen unterschieden:

1. Die N Individuen, welche die Population bilden, werden in eine Rangreihe gebracht.
2. Aus der Population vom Umfang N wird eine Zufallsstichprobe vom Umfang n gezogen. Die Individuen der Stichprobe werden in eine Rangreihe gebracht.

In beiden Fällen sollen Parameter der Verteilung der Rangzahlen eines Individuums i (z.B. der wahre Wert) oder Kennzahlen des Tests (z.B. die Reliabilität) geschätzt werden, wobei erwartungstreue Schätzungen der Populationsparameter gewünscht werden. Die Verteilung der N! Rangordnungen ist in beiden Fällen unbekannt. Es wird angenommen, daß k Rater vorhanden sind, welche unabhängig voneinander die Individuen beurteilen, so daß k Rangreihen entstehen. Die Rater werden als *homogen* vorausgesetzt, was bedeutet, daß diese einem Individuum i gleiche wahre Werte und gleiche Varianzen der Rangzahlen zuweisen. Die Verteilungen der Rangordnungen der einzelnen Rater müssen nicht gleich sein. Sind $R_i^1, R_i^2, \ldots, R_i^k$ die Rangzahlen, welche die k Rater dem Individuum i zuweisen, so ist bei homogenen Ratern

$$E_a[R_i^1] = E_a[R_i^2] = \ldots = E_a[R_i^k] = t_i \quad , \text{ und}$$

$$Var_a[R_i^1] = Var_a[R_i^2] = \ldots = Var_a[R_i^k] = Var_a[R_i] \quad .$$

Damit sind homogene Rater parallele Rater mit der zusätzlichen Forderung der Varianzhomogenität. Die Voraussetzung lokal unkorrelierter Rater statt unabhängiger Rater reicht für die Schätzun-

gen aus. Wie aber in 1.2. schon bemerkt wurde, sind Rater, die lokal unkorreliert voneinander urteilen, aber nicht unabhängig sind, schwer vorstellbar. Deshalb wird hier weiterhin von unabhängigen Ratern gesprochen. Es sei aber darauf hingewiesen, daß die Voraussetzung der lokalen Unkorreliertheit ausreicht.
Das Schätzen der gewünschten Parameter im Fall 1 entspricht weitgehend dem Vorgehen in der klassischen Statistik, es wurde von Schulman nicht erwähnt.

Im Fall 2 jedoch kommen die Besonderheiten des ordinalen Testmodells zum Tragen. Die Vorschläge Schulmans (1976, 1978) zur Schätzung des wahren Wertes, der Varianz der Rangzahlen und der Reliabilität wurden z.T. übernommen oder neu gestaltet. So wurde z.B. die etwas umständliche Schätzung der Reliabilität nicht dargestellt, ein Konfidenzintervall für den wahren Wert wurde hinzugefügt.

Nimmt man an, daß aus der Population n Individuen zufällig ausgewählt werden, $1 < n \leq N$, $2 < N$, so hat man in diesem Modell zusätzlich zu den Zufallsvariablen $(R_1,\ldots,R_N)$, welche die Zuweisung der Rangzahlen zu den N Individuen der Population bedeuten, noch die Zufallsvariablen $(L_1,\ldots,L_n)$ zu beachten. Z.B. ist L_1 die Zufallsvariable, welche das zuerst ausgewählte Individuum angibt. Die Zufallsvariablen $(R_1,\ldots,R_N)$ werden als stochastisch unabhängig von $(L_1,\ldots,L_n)$ vorausgesetzt, d.h. die Zuweisung der Rangzahlen in der Population ist unabhängig von der Auswahl der Stichprobe.
Weiterhin wird angenommen, daß alle möglichen Zufallsstichproben die gleiche Wahrscheinlichkeit haben, ausgewählt zu werden,

$$P((L_1,\ldots,L_n) = (l_1,\ldots,l_n)) = \frac{N!}{(N-n)!} .$$

Die n ausgewählten Individuen werden in eine Rangreihe gebracht. Es gibt n! mögliche Rangreihen. Die Rangzahl des Individuums l_i in der Stichprobe sei mit $R^*_{l_i}$ bezeichnet. Die Wahrscheinlichkeit für eine mögliche Anordnung $(k_1,\ldots,k_n)$ der n Individuen der Stichprobe, unter der Bedingung, daß es die Individuen $(l_1,\ldots,l_n)$ sind, welche aus der Population ausgewählt wurden, soll mit $q_{k_1,\ldots,k_n|l_1,\ldots,l_n}$ bezeichnet werden.

Es ist

$$q_{k_1,\ldots,k_n|l_1,\ldots,l_n} = P(\{R^*_{l_1} = k_1,\ldots,R^*_{l_n} = k_n\}|\{L_1 = l_1,\ldots,L_n = l_n\})$$

$$= \frac{P(\{R^*_{l_1} = k_1,\ldots,R^*_{l_n} = k_n\}\cap\{L_1 = l_1,\ldots,L_n = l_n\})}{P(L_1 = l_1,\ldots,L_n = l_n)} .$$

Die Verteilung der Rangordnungen in der Stichprobe wird durch die Verteilung der Rangordnungen in der Population bestimmt. Denn gilt für die Rangzahlen zweier Individuen i_1, i_2 bezüglich der Population $j_{i_1} < j_{i_2}$, so soll für dieselben Individuen auch bezüglich der Stichprobe gelten $k_{i_1} < k_{i_2}$. Dann ist

$$q_{k_1,\ldots,k_n|l_1,\ldots,l_n} = \sum_{*} p_{j_1,\ldots,j_N} .$$

Dabei bedeutet "$*$": alle $(j_1,\ldots,j_N)$ mit

$$k_1 = \text{Rang}\,(j_{l_1}),\ldots,k_n = \text{Rang}\,(j_{l_n}) ,$$

wobei die Ränge in Bezug auf die Stichprobe gebildet werden.

Soll für ein bestimmtes Individuum i der Population z.B. der wahre Wert t_i oder die Varianz der Rangzahlen $\text{Var}_a[R_i]$ aus den Stichprobendaten geschätzt werden, so geht man wie folgt vor: Das Individuum i wird aus der Population ausgewählt. Aus den verbleibenden (N-1) Individuen wird dann eine Zufallsstichprobe vom Umfang (n-1) gezogen, so daß insgesamt eine Stichprobe vom Umfang n vorliegt: $(i,l_1,\ldots,l_{n-1})$. Alles oben über die Zufallsstichprobe vom Umfang n Gesagte, gilt dann für die Zufallsstichprobe vom Umfang (n-1). Es handelt sich jetzt um eine bedingte Betrachtung. Die Bedingung besteht darin, daß das n-te Element der Stichprobe das Individuum i ist: $l_n = i$.

Erwartungswerte können sich auf die zufällige Auswahl der Stichprobe beziehen (E_z) und auf die Zuweisung der Rangzahlen in der Population (E_a). Die Schätzungen sollen erwartungstreu sein bezüglich der Zufallsauswahl der Stichprobe und bezüglich der Zuweisung der Rangzahlen.

2.1. Schätzen des wahren Wertes eines festen Individuums

2.1.1. Punktschätzung

2.1.1.1. Die N Individuen der Population werden in eine Rangreihe gebracht

Es wird angenommen, daß k homogene Rater unabhängig voneinander die N Individuen in Rangreihen bringen. Ziel ist es, für ein festes Individuum i dessen wahren Wert t_i zu schätzen.

Eine solche Schätzung des wahren Wertes t_i ist

$$\hat{T}_i = \frac{1}{k} \sum_{j=1}^{k} R_i^j \quad .$$

R_i^j ist die Rangzahl des Individuums i in der Rangordnung des Raters j. Damit wird der wahre Wert eines Individuums durch den Rangmittelwert geschätzt. Für k=1 ist dieser Wert gleich der Rangzahl. Die Schätzfunktion $\hat{T}_i$ ist erwartungstreu bezüglich der Rangordnungen:

$$E_a[\hat{T}_i] = \frac{1}{k} \sum_{j=1}^{k} E_a[R_i^j] = \frac{1}{k} \sum_{j=1}^{k} t_i = \frac{1}{k} \, kt_i = t_i \quad .$$

Bei steigender Anzahl von Ratern wird $Var_a[\hat{T}_i]$ kleiner:

$$Var_a[\hat{T}_i] = \frac{1}{k^2} \sum_{j=1}^{k} Var_a[R_i^j] = \frac{1}{k} \, Var_a[R_i] \quad .$$

2.1.1.2. Die n Individuen der Stichprobe werden in eine Rangreihe gebracht

Um den wahren Wert t_i des Individuums i zu schätzen, werden die n Individuen der Stichprobe $(i,l_1,\ldots,l_{n-1})$ von k Ratern in Rangreihen gebracht. Der wahre Wert des Individuums i in Bezug auf die Stichprobe ist nicht von Interesse, da er von der Zusammensetzung der Stichprobe abhängig ist. Es soll der wahre Wert in Bezug auf die Population geschätzt werden, d.h. aus der Rangzahl $R_i^{(n-1)}$ eines Individuums i in Bezug auf die Stichprobe vom Umfang (n-1) soll der wahre Wert t_i in Bezug auf die Population geschätzt werden. Gewünscht ist eine Schätzfunktion $\hat{T}_i^{(n-1)}$, die

erwartungstreu ist in Bezug auf die zufällige Auswahl der Stichprobe und bezüglich der Wahrscheinlichkeitsverteilung der Rangordnungen in der Population:

$$E_{az}[\hat{T}_i^{(n-1)}] = E_a[E_z[\hat{T}_i^{(n-1)} \mid R]] = t_i .$$

Dieses leistet die Schätzfunktion:

$$\hat{T}_i^{(n-1)} = \frac{(N-1)R_i^{(n-1)} - (N-n)}{n-1} ,$$

wobei $R_i^{(n-1)}$ die Rangzahl des ausgewählten Individuums i in Bezug auf die Stichprobe ist. Die Berechnung dieser Schätzfunktion bedeutet eine Hochrechnung von der Stichprobe auf die Population. Dabei werden die extremen Rangplätze, d.h. Rang 1 und Rang n, wieder zu extremen Rangplätzen (Rang 1 and Rang N). Der Umfang N der Population muß bekannt sein.

Um die Erwartungstreue von $\hat{T}_i^{(n-1)}$ zu zeigen, wird die Zufallsvariable

$Y_i^{(n-1)} = R_i^{(n-1)} - 1$ = "Anzahl der Individuen, die eine kleinere Rangzahl haben als Individuum i (in der ausgewählten Stichprobe vom Umfang (n-1))"

definiert.

Stellt man sich zunächst eine feste Populationsanordnung $j_1,\ldots,j_N$ und ein festes Individuum i vor, so ist $Y_i^{(n-1)}$ hypergeometrisch verteilt.

Die hypergeometrische Verteilung läßt sich am Urnenmodell folgendermaßen interpretieren: Eine Urne enthalte N Kugeln, davon S rote und N-S weiße. Werden n Kugeln zufällig und ohne Zurücklegen aus der Urne entnommen, so gibt P(X=s) die Wahrscheinlichkeit dafür an, daß sich unter den entnommenen Kugeln genau s rote befinden. Es ist

$$E[X] = \frac{S \cdot n}{N} \quad \text{und} \quad \mathrm{Var}[X] = \frac{S\,n\,(N-S)(N-n)}{N^2(N-1)} .$$

Identifiziert man die Anzahl der Individuen, die einen kleineren Rang haben als Individuum i mit den roten Kugeln in der Urne

$(S \hat{=} j_i - 1)$, die Anzahl der Individuen, die einen größeren Rang haben als Individuum i mit den weißen Kugeln in der Urne $(N-S \hat{=} (N-1)-(j_i-1) = N-j_i)$, so hat man ein Urnenmodell mit insgesamt (N-1) Kugeln und $Y_i^{(n-1)}$ = "Anzahl der roten Kugeln in der Stichprobe vom Umfang (n-1)" ist hypergeometrisch verteilt.

Es ist

$$E_z[Y_i^{(n-1)}] = (n-1)\,\frac{j_i-1}{N-1}\,,$$

$$\mathrm{Var}_z[Y_i^{(n-1)}] = \frac{(N-n)(n-1)}{(N-2)(N-1)^2}\,(j_i-1)(N-j_i)\,.$$

Damit erhält man

$$E_a[E_z[\hat{T}_i^{(n-1)}\,|\,R]] =$$

$$= E_a[E_z[\frac{N-1}{n-1}R_i^{(n-1)} - \frac{N-n}{N-1}\,|\,R]] = E_a[E_z[\frac{N-1}{n-1}\,(R_i^{(n-1)}-1)+1\,|\,R]]$$

$$= E_a[E_z[\frac{N-1}{n-1}\,Y_i^{(n-1)}+1\,|\,R]] = E_a[(n-1)\,\frac{N-1}{n-1}\,\frac{R_i-1}{N-1}+1]$$

$$= E_a[R_i] = t_i\,.$$

Damit ist gezeigt, daß $\hat{T}_i^{(n-1)} = \dfrac{(N-1)R_i^{(n-1)}-(N-n)}{n-1}$ erwartungstreu bezüglich der Zufallsauswahl der Stichprobe und der Verteilung der Rangordnungen in der Population ist.

Will man statt für t_i eine Schätzfunktion für $\frac{t_i}{N}$ angeben, so erhält man

$$\frac{\hat{T}_i^{(n-1)}}{N} = \frac{N-1}{N}\,\frac{R_i^{(n-1)}-1}{n-1} + \frac{1}{N}\,.$$

Für sehr großes N ergibt sich näherungsweise

$$\frac{\hat{T}_i^{(n-1)}}{N} \approx \frac{R_i^{(n-1)}-1}{n-1}\,.$$

Diese Schätzfunktion ist unabhängig von N.

Die Varianz der Schätzfunktion $\hat{T}_i^{(n-1)}$ ist für steigenden Stichprobenumfang monoton fallend:

$$Var_{az}[\hat{T}_i^{(n-1)}] = Var_{az}[\frac{N-1}{n-1}(R_i^{(n-1)}-1)+1] = (\frac{N-1}{n-1})^2 Var_{az}[R_i^{(n-1)}-1]$$

$$= (\frac{N-1}{n-1})^2 Var_{az}[Y_i^{(n-1)}] .$$

Mit dem Verschiebungssatz und den oben erhaltenen Ergebnissen bezüglich $Y_i^{(n-1)}$ erhält man für $Var_{az}[Y_i^{(n-1)}]$:

$$Var_{az}[Y_i^{(n-1)}] = E_{az}[Y_i^{(n-1)2}] - (E_{az}[Y_i^{(n-1)}])^2$$

$$= E_a[E_z[Y_i^{(n-1)2}|R]] - (E_a[E_z[Y_i^{(n-1)}|R]])^2$$

$$= E_a[Var_z[Y_i^{(n-1)}|R] + (E_z[Y_i^{(n-1)}|R])^2] - (E_a[\frac{N-1}{n-1}(R_i-1)])^2$$

$$= E_a[\frac{(N-n)(n-1)}{(N-2)(N-1)^2}(R_i-1)(N-R_i) + (\frac{n-1}{N-1})^2 (R_i-1)^2] -$$

$$-(\frac{n-1}{N-1})^2 (t_i-1)^2 .$$

Damit ist

$$Var_{az}[\hat{T}_i^{(n-1)}] = (\frac{N-1}{n-1})^2 \cdot \frac{(N-n)(n-1)}{(N-2)(N-1)^2} E_a[(R_i-1)(N-R_i)] + E_a[(R_i-1)^2] -$$

$$-(t_i-1)^2$$

$$= \frac{N-n}{(N-2)(n-1)} E_a[NR_i - R_i^2 - N + R_i] + Var_a[R_i-1]$$

$$= \frac{N-n}{(N-2)(n-1)} E_a[(\frac{N-1}{2})^2 - (R_i - \frac{N+1}{2})^2] + Var_a[E_i]$$

$$= \frac{N-n}{(N-2)(n-1)}[(\frac{N-1}{2})^2 - Var_a[E_i] - (t_i - \frac{N+1}{2})^2] + Var_a[E_i] .$$

Für steigenden Stichprobenumfang wird die Varianz der Schätzfunktion $\hat{T}_i^{(n-1)}$ kleiner, bis für n=N das Minimum $Var_a[E_i]$ erreicht ist.

Die Varianz setzt sich zusammen aus einem Anteil, der durch das Schätzen aus der Stichprobe entsteht, und aus der Varianz der Rangzahlen des Individuums i, $Var_a[E_i]$. Für n=N fällt der Anteil, welcher durch das Schätzen entsteht, weg.

Stehen statt nur eines Raters k homogene unabhängige Rater zur Beurteilung der n Individuen der Stichprobe zur Verfügung, so kann $\hat{T}_i^{(n-1)}$ für jede der k Rangreihen berechnet werden. Da $\hat{T}_i^{(n-1)}$ erwartungstreu für t_i ist, ist es auch

$$\frac{1}{k} \sum_{j=1}^{k} \hat{T}_{ij}^{(n-1)} = \bar{\hat{T}}_i^{(n-1)} .$$

Es ist

$$\bar{\hat{T}}_i^{(n-1)} = \frac{(N-1)\bar{R}_i^{(n-1)} - (N-n)}{n-1}$$

und

$$Var_{az}[\bar{\hat{T}}_i^{(n-1)}] = \frac{1}{k} Var_{az}[\hat{T}_i^{(n-1)}] .$$

2.1.2. Konfidenzintervall für den wahren Wert

2.1.2.1. Die N Individuen der Population werden in eine Rangreihe gebracht

Wenn keinerlei Annahmen bezüglich der Wahrscheinlichkeitsverteilung bestehen, wird ein Konfidenzintervall für den wahren Wert mit Hilfe der Tschebyscheffschen Ungleichung berechnet.

Angenommen, es liegen k Rangreihen von k homogenen unabhängigen Urteilern vor. Das Individuum i erhält dann die mittlere Rangzahl

$$\bar{R}_i = \frac{1}{k} \sum_{j=1}^{k} R_i^j$$

zugewiesen mit

$$Var_a[\bar{R}_i] = \frac{1}{k^2} Var_a[\sum_{j=1}^{k} R_i^j] = \frac{1}{k^2} k \, Var_a[R_i] = \frac{1}{k} Var_a[E_i] .$$

Mit $P(|\bar{R}_i - t_i| \geq \varepsilon \sqrt{Var_a[\bar{R}_i]}) \leq \frac{1}{\varepsilon^2}$

oder

$$P(|\bar{R}_i - t_i| \leq \varepsilon \sqrt{Var_a[\bar{R}_i]}) \geq 1 - \frac{1}{\varepsilon^2}$$

lautet ein Konfidenzintervall (Niveau $1 - \frac{1}{\varepsilon^2}$):

$$\bar{R}_i - \varepsilon \sqrt{\frac{1}{k} Var_a[E_i]} \leq t_i \leq \bar{R}_i + \varepsilon \sqrt{\frac{1}{k} Var_a[E_i]} .$$

Ist die Varianz $Var_a[E_i]$ unbekannt, so kann sie bei k Urteilern mit Hilfe der Schätzfunktion

$$\frac{1}{k-1} \sum_{j=1}^{k} (R_i^j - \bar{R}_i)^2$$

geschätzt werden (siehe nächster Abschnitt). Es handelt sich dann allerdings nicht mehr um ein Konfidenzintervall.

Mit der Annahme, daß alle Individuen gleiche Varianzen der Rangzahlen haben, kann die Schätzung über die Gruppenfehlervarianz erfolgen:

$$Var_g[R] = Var_g[T] + Var_g[E],$$

$$1 = \frac{Var_g[T]}{Var_g[R]} + \frac{Var_g[E]}{Var_g[R]} = \rho^2[R,T] + \frac{Var_g[E]}{Var_g[R]} .$$

Damit ergibt sich für $Var_g[E]$:

$$Var_g[E] = (1-\rho^2[R,T]) Var_g[R] = (1-\rho^2(R,T)) \frac{N^2-1}{12} .$$

Dabei muß $\rho^2[R,T]$ bekannt sein, oder es wird mit $\hat{\rho}^2$ geschätzt. Die Annahme gleicher Varianzen ist allerdings sehr unrealistisch.

2.1.2.2. Die n Individuen der Stichprobe werden in eine Rangreihe gebracht

Mit der Tschebyscheffschen Ungleichung für die im letzten Abschnitt abgeleitete Schätzfunktion $\hat{T}_i^{(n-1)}$ für den wahren Wert t_i erhält

man das Konfidenzintervall zum Niveau $1-\frac{1}{\varepsilon^2}$:

$$|\hat{T}_i^{(n-1)}-t_i| \leq \varepsilon \sqrt{Var_{az}[\hat{T}_i^{(n-1)}]} .$$

Die Auflösung der Ungleichung nach t_i ist hier nicht so einfach wie in Abschnitt 2.1.2.1., da $Var_{az}[\hat{T}_i^{(n-1)}]$ von t_i abhängig ist. In der folgenden Rechnung wird der Einfachheit halber statt der Ungleichung die Gleichung betrachtet. Man erhält so die obere (und die untere) Grenze des Konfidenzintervalls. Quadrieren ergibt

$$(\hat{T}_i^{(n-1)}-t_i)^2 = \varepsilon^2 Var_{az}[\hat{T}_i^{(n-1)}] .$$

Damit die Rechnung übersichtlich bleibt, wird statt für t_i ein Konfidenzintervall für $(t_i - \frac{N+1}{2})$ berechnet.

$$(\hat{T}_i^{(n-1)}- \tfrac{N+1}{2})^2 + (t_i - \tfrac{N+1}{2})^2 - 2(t_i - \tfrac{N+1}{2})(\hat{T}_i^{(n-1)}- \tfrac{N+1}{2}) = \varepsilon^2 Var_{az}[\hat{T}_i^{(n-1)}] .$$

Mit

$$Var_{az}[\hat{T}_i^{(n-1)}] = \frac{N-n}{(N-2)(n-1)} \{ (\tfrac{N-1}{2})^2 - (t_i - \tfrac{N+1}{2})^2 - Var_a[E_i] \} + Var_a[E_i]$$

ergibt sich die Umformung

$$(t_i - \tfrac{N+1}{2})^2 (1+\varepsilon^2 A) - 2(t_i - \tfrac{N+1}{2})(\hat{T}_i - \tfrac{N+1}{2}) + (\hat{T}_i - \tfrac{N+1}{2})^2 - \varepsilon^2 A (\tfrac{N-1}{2})^2 +$$

$$+ \varepsilon^2 Var_a[E_i](A-1) = 0 ,$$

wobei $\frac{N-n}{(n-1)(N-2)} = A$ gesetzt und der Einfachheit halber statt $\hat{T}_i^{(n-1)}$ nur $\hat{T}_i$ geschrieben wurde.

Auflösen der quadratischen Gleichung ergibt

$$(t_i - \tfrac{N+1}{2})_{1,2} = \frac{\hat{T}_i - \frac{N+1}{2}}{1+\varepsilon^2 A} \pm \frac{\varepsilon}{\sqrt{1+\varepsilon^2 A}} \cdot \sqrt{(\hat{T}_i - \tfrac{N+1}{2})^2 \frac{-A}{1+\varepsilon^2 A} + A(\tfrac{N-1}{2})^2 +}$$

$$\overline{+Var_a[E_i](1-A)} .$$

Zur Vereinfachung wird gesetzt

$$B = \frac{1}{\sqrt{1+\varepsilon^2 A}} \sqrt{(\hat{T}_i - \frac{N+1}{2})^2 \frac{-A}{1+\varepsilon^2 A} + A(\frac{N-1}{2})^2 + Var_a[E_i](1-A)} \quad .$$

Damit ist

$$(t_i - \frac{N+1}{2})_{1,2} = \frac{\hat{T}_i - \frac{N+1}{2}}{1+\varepsilon^2 A} \pm \varepsilon B \quad .$$

Ein Konfidenzintervall zum Niveau $1-\frac{1}{\varepsilon^2}$ lautet damit

$$\frac{\hat{T}_i - \frac{N+1}{2}}{1+\varepsilon^2 A} - \varepsilon B \leq t_i - \frac{N+1}{2} \leq \frac{\hat{T}_i - \frac{N+1}{2}}{1+\varepsilon^2 A} + \varepsilon B \quad .$$

Die Rechnung wurde für nur eine genannte Rangreihe durchgeführt. Wurde $\bar{\hat{T}}_i$ mit Hilfe von k homogenen Ratern gewonnen, so ist $Var_{az}[\bar{\hat{T}}_i^{(n-1)}] = \frac{1}{k} Var_{az}[\hat{T}_i^{(n-1)}]$. Das angegebene Konfidenzintervall ändert sich insofern, als statt $\hat{T}_i$ dann $\bar{\hat{T}}_i$ steht und statt des Faktors ε^2 jetzt ε^2/k.
Berechnet man das Konfidenzintervall für $\frac{t_i - \frac{N+1}{2}}{N}$ statt für $(t_i - \frac{N+1}{2})$ und nimmt man weiterhin an, daß N sehr groß ist im Vergleich zu n, so ergibt sich näherungsweise das Intervall

$$\frac{\frac{\hat{T}_i}{N} - \frac{1}{2}}{1+\varepsilon^2 A} - \varepsilon \cdot D \leq \frac{t_i}{N} - \frac{1}{2} \leq \frac{\frac{\hat{T}_i}{N} - \frac{1}{2}}{1+\varepsilon^2 A} + \varepsilon D$$

mit $A \cong \frac{1}{n-1}$ und

$$D = \frac{1}{\sqrt{1+\varepsilon^2 A}} \sqrt{(\frac{\hat{T}_i}{N} - \frac{1}{2})^2 \frac{-A}{1+\varepsilon^2 A} + \frac{A}{4}} \quad .$$

Das Intervall hängt nicht mehr von $Var_a[E_i]$ ab.

2.2. Schätzen der Varianz der Rangzahlen des Individuums i

2.2.1. Die N Individuen der Population werden in eine Rangreihe gebracht

Es wird angenommen, daß k Rangreihen der N Individuen vorliegen, welche k homogene Rater unabhängig voneinander erstellten. $R_i^1,\ldots,R_i^k$ seien die Rangzahlen, welche die k Rater dem Individuum i zuweisen. Es ist $E_a[R_i^1] = \ldots = E_a[R_i^k] = t_i$ und $Var_a[R_i^1] = \ldots = Var_a[R_i^k] = Var_a[R_i]$.

Die Varianz $Var_a[R_i]$ der Rangzahlen des Individuums i soll geschätzt werden. Dieses kann mit Hilfe der empirischen Varianz geschehen:

$$\hat{\sigma}^2(R_i) = \frac{1}{k-1} \sum_{j=1}^{k} (R_i^j - \bar{R}_i)^2 \ .$$

$\hat{\sigma}^2(R_i)$ ist eine erwartungstreue Schätzfunktion für $Var_a[R_i]$ bezüglich der Rangordnungen:

$$E_a[\hat{\sigma}^2] = \frac{1}{k-1} \sum_{j=1}^{k} E_a[((R_i^j - t_i) - (\bar{R}_i - t_i))^2] =$$

$$= \frac{1}{k-1} \sum_{j=1}^{k} (Var_a[R_i] + Var_a[\bar{R}_i] - 2E_a((R_i^j - t_i)(\bar{R}_i - t_i))$$

$$= \frac{1}{k-1} (k\ Var_a[R_i] + Var_a[R_i] - 2Var_a[R_i])$$

$$= Var_a[R_i] \ .$$

Für 2 Populationsanordnungen, welche von 2 unabhängigen homogenen Ratern j, j' genannt wurden und welche zu den Rangzahlen R_i^j und $R_i^{j'}$ für Individuum i führen, kann $Var_a[R_i]$ auch wie folgt geschrieben werden:

$$Var_a[R_i] = E_a\left[\frac{(R_i^j - R_i^{j'})^2}{2}\right] = E_a\left[\frac{D_{ijj'}^2}{2}\right] \ ,$$

wobei der Erwartungswert E_a bezüglich der bivariaten Verteilung, nach der beide Rater urteilen, gebildet wird.

Bei 2 homogenen Ratern kann also $Var_a[R_i]$ auch erwartungstreu geschätzt werden mit:

$$\hat{\sigma}^2(R_i) = \frac{(R_i^j - R_i^{j'})^2}{2} = \frac{D_{ijj'}^2}{2} .$$

Bei k homogenen Ratern wird für alle $\binom{k}{2}$ Paare von Ratern $\frac{D_{ijj'}^2}{2}$ berechnet und dann der Mittelwert gebildet:

$$\hat{\sigma}^2[R_i] = \frac{1}{\binom{k}{2}} \sum_{j<j'} \frac{D_{ijj'}^2}{2} .$$

Da

$$\frac{1}{\binom{k}{2}} \sum_{j<j'} \frac{D_{ijj'}^2}{2} = \frac{1}{k-1} \sum_{j=1}^{k} (R_i^j - \bar{R}_i)^2$$

gilt, ist auch diese Schätzung erwartungstreu.

2.2.2. Die n Individuen der Stichprobe werden in eine Rangreihe gebracht

Es wird, wie in Abschnitt 2.1.1.2., angenommen, daß aus der Population zuerst das Individuum i ausgewählt wird. Anschließend wird eine Zufallsstichprobe vom Umfang (n-1) gezogen, so daß insgesamt eine Stichprobe vom Umfang n vorliegt.

Es wird weiterhin angenommen, daß zwei homogene unabhängige Rater j, j' vorliegen, welche den Individuen Rangzahlen zuweisen. Die Rangzahlen bezüglich der Population seien R_{ij} bzw. $R_{ij'}$, diejenigen bezüglich der Stichprobe $R_{ij}^{(n-1)}$ und $R_{ij'}^{(n-1)}$.

Die Varianz $Var_a[R_i^{(n-1)}]$ des Individuums i in Bezug auf die Stichprobe kann (analog zu 2.2.1.) erwartungstreu geschätzt werden mit

$$\hat{\sigma}^2(R_i^{(n-1)}) = \frac{1}{2} D_{ijj'}^{(n-1)2} = \frac{(R_{ij}^{(n-1)} - R_{ij'}^{(n-1)})^2}{2} .$$

$\hat{\sigma}^2(R_i^{(n-1)})$ ist keine erwartungstreue Schätzung der Varianz $Var_a[R_i]$.

Um eine erwartungstreue Schätzung für die Varianz der Rangzahlen des Individuums i in Bezug auf die Population abzuleiten, wird

$$E_{az}\left[\frac{D_{ijj'}^{(n-1)2}}{2}\right] = E_a\left[E_z\left[\frac{D_{ijj'}^{(n-1)2}}{2} \mid R\right]\right]$$

berechnet.

Um die Berechnung zu erleichtern, seien einige Größen definiert. Es wird zunächst angenommen, daß 2 feste Populationsanordnungen, welche von den Ratern j und j' genannt wurden, vorliegen.

Für diese beiden Anordnungen wird definiert:

$A_{1ijj'}$: Anzahl der Individuen, welche in beiden Rangordnungen einen kleineren Rang haben als Individuum i,

$A_{4ijj'}$: Anzahl der Individuen, welche in beiden Rangordnungen einen größeren Rang haben als Individuum i,

$A_{2ijj'}$: Anzahl der Individuen, welche in Rangordnung j' einen größeren, in Rangordnung j einen kleineren Rang haben als Individuum i,

$A_{3ijj'}$: Anzahl der Individuen, welche in Rangordnung j' einen kleineren, in Rangordnung j einen größeren Rang haben als Individuum i.

Es ist

$$A_{1ijj'}+A_{2ijj'}+A_{3ijj'}+A_{4ijj'} = N-1,\ A_{1ijj'}+A_{2ijj'} = R_{ij}-1,$$

$$A_{1ijj'}+A_{3ijj'} = R_{ij'}-1 .$$

Es werde bezeichnet:

$$B_{ijj'} = A_{2ijj'}+A_{3ijj'} = R_{ij}+R_{ij'}-2A_{1ijj'}-2 .$$

$B_{ijj'}$ ist ein Maß, welches die Veränderung der Position des

Individuums i in den beiden Rangordnungen widerspiegelt.

Für die Stichprobe hat man entsprechend:

$$A_{1ijj'}^{(n-1)} + A_{2ijj'}^{(n-1)} + A_{3ijj'}^{(n-1)} + A_{4ijj'}^{(n-1)} = n-1 \ ,$$

$$A_{1ijj'}^{(n-1)} + A_{2ijj'}^{(n-1)} = R_{ij}^{(n-1}-1, \ A_{1ijj'}^{(n-1)} + A_{3ijj'}^{(n-1)} = R_{ij'}^{(n-1)}-1$$

und

$$B_{ijj'}^{(n-1)} = A_{2ijj'}^{(n-1)} + A_{3ijj'}^{(n-1)} \ .$$

Damit ist

$$D_{ijj'}^{(n-1)2} = (R_{ij}^{(n-1)}-R_{ij'}^{(n-1)})^2 = (A_{1ijj'}^{(n-1)} + A_{2ijj'}^{(n-1)} - A_{1ijj'}^{(n-1)} - A_{3ijj'}^{(n-1)})^2$$

$$= (A_{2ijj'}^{(n-1)} - A_{3ijj'}^{(n-1)})^2$$

$$E_z[D_{ijj'}^{(n-1)2}|R] = E_z[(A_{2ijj'}^{(n-1)} - A_{3ijj'}^{(n-1)})^2|R] =$$

$$= Var_z[(A_{2ijj'}^{(n-1)} - A_{3ijj'}^{(n-1)})|R] +$$

$$+ (E_z[A_{2ijj'}^{(n-1)} - A_{3ijj'}^{(n-1)})|R])^2$$

$$= Var_z[A_{2ijj'}^{(n-1)}|R] + Var_z[A_{3ijj'}^{(n-1)}|R] -$$

$$- 2Cov_z[A_{2ijj'}^{(n-1)}, A_{3ijj'}^{(n-1)}|R] +$$

$$+ (E_z[(A_{2ijj'}^{(n-1)} - A_{3ijj'}^{(n-1)})|R])^2 \ .$$

Mit

$$2Cov_z[A_{2ijj'}^{(n-1)}, A_{3ijj'}^{(n-1)}|R] = Var_z[(A_{2ijj'}^{(n-1)} + A_{3ijj'}^{(n-1)})|R] -$$

$$- Var_z[A_{2ijj'}^{(n-1)}|R] - Var_z[A_{3ijj'}^{(n-1)}|R]$$

erhält man

$$E_z[D_{ijj'}^{(n-1)2}|R] = 2Var_z[A_{2ijj'}^{(n-1)}|R] + 2Var_z[A_{3ijj'}^{(n-1)}|R] -$$

$$- \mathrm{Var}_z[(A_{2ijj'}^{(n-1)} + A_{3ijj'}^{(n-1)})|R] +$$

$$+ (E_z[(A_{2ijj'}^{(n-1)} - A_{3ijj'}^{(n-1)})|R])^2 .$$

$A_{2ijj'}^{(n-1)}$, $A_{3ijj'}^{(n-1)}$ und $(A_{2ijj'}^{(n-1)} + A_{3ijj'}^{(n-1)})$ sind hypergeometrisch verteilt. Man stellt sich dabei die folgenden drei Urnenmodelle vor:

Die Urne enthält $A_{2ijj'}$ rote und $(A_{1ijj'}+A_{3ijj'}+A_{4ijj'})$ schwarze Kugeln,
die Urne enthält $A_{3ijj'}$ rote und $(A_{1ijj'}+A_{2ijj'}+A_{4ijj'})$ schwarze Kugeln,
die Urne enthält $(A_{2ijj'}+A_{3ijj'})$ rote und $(A_{1ijj'}+A_{4ijj'})$ schwarze Kugeln.

Für jedes Modell ist die "Anzahl der roten Kugeln in der Stichprobe vom Umfang (n-1)" hypergeometrisch verteilt mit

$$E_z[A_{2ijj'}^{(n-1)}|R] = (n-1)\,\frac{A_{2ijj'}}{N-1}\,,\quad E_z[A_{3ijj'}^{(n-1)}|R] = (n-1)\,\frac{A_{3ijj'}}{N-1}\,,$$

$$\mathrm{Var}_z[A_{2ijj'}^{(n-1)},] = (n-1)\,\frac{N-n}{N-2}\,\frac{A_{2ijj'}}{N-1}\cdot\frac{N-1-A_{2ijj'}}{N-1}\,,$$

$$\mathrm{Var}_z[A_{3ijj'}^{(n-1)}|R] = (n-1)\,\frac{N-n}{N-2}\,\frac{A_{3ijj'}}{N-1}\,\frac{N-1-A_{3ijj'}}{N-1}\,,$$

$$\mathrm{Var}_z[(A_{2ijj'}^{(n-1)} + A_{3ijj'}^{(n-1)})|R] = (n-1)\,\frac{N-n}{N-2}\,\frac{A_{2ijj'}+A_{3ijj'}}{N-1}\cdot$$

$$\frac{N-1-A_{2ijj'}-A_{3ijj'}}{N-1}\,.$$

Dieses wird in die Formel für $E_z[D_{ijj'}^{(n-1)2}|R]$ eingesetzt:

$$E_z[D_{ijj'}^{(n-1)2}|R] = (n-1)\,\frac{N-n}{N-2}\,(2\,\frac{A_{2ijj'}}{N-1}\,\frac{N-1-A_{2ijj'}}{N-1} +$$

$$+ 2\,\frac{A_{3ijj'}}{N-1}\,\frac{N-1-A_{3ijj'}}{N-1} -$$

$$- \frac{A_{2ijj'}+A_{3ijj'}}{N-1}\,\frac{N-1-A_{2ijj'}-A_{3ijj'}}{N-1})\, +$$

$$+ (\frac{n-1}{N-1})^2\,(A_{2ijj'}-A_{3ijj'})^2\,.$$

Nach geeignetem Zusammenfassen ergibt sich

$$E_z[D_{ijj'}^{(n-1)2}|R] = \frac{n-1}{N-1}\frac{N-n}{N-2}(A_{2ijj'}+A_{3ijj'})^2 + \frac{n-1}{N-1}\frac{n-2}{N-2}(A_{2ijj'}-A_{3ijj'})^2 .$$

Bildet man den Erwartungswert bezüglich der Rangordnungen, so erhält man

$$E_a[E_z[\frac{1}{2}D_{ijj'}^{(n-1)2}|R]] = \frac{n-1}{N-1}\frac{N-n}{N-2}\frac{1}{2}E_a[A_{2ijj'}+A_{3ijj'}]^2 + \frac{n-1}{N-1}\frac{n-2}{N-2}\mathrm{Var}_a[R_i] .$$

Man sieht, daß $D_{ijj'}^{(n-1)2}$ keine erwartungstreue Schätzung für $\mathrm{Var}_a[R_i]$ ist. Eine solche erhält man durch Umformung

$$\begin{aligned}\mathrm{Var}_a[R_i] &= \frac{(N-1)(N-2)}{(n-1)(n-2)}E_a[E_z[\frac{D_{ijj'}^{(N-1)2}}{2}|R]] - \\ &\quad - \frac{N-n}{n-2}E_a[\frac{A_{2ijj'}+A_{3ijj'}}{2}]^2 \\ &= \frac{(N-1)(N-2)}{(n-1)(n-2)}E_a[E_z[\frac{D_{ijj'}^{(n-1)2}}{2}|R]] - \\ &\quad - \frac{(N-n)(N-1)}{(n-2)(n-1)}E_a[E_z[\frac{A_{2ijj'}^{(n-1)}+A_{3ijj'}^{(n-1)}}{2}|R]] .\end{aligned}$$

Mit

$$B_{ijj'}^{(n-1)} = A_{2ijj'}^{(n-1)} + A_{3ijj'}^{(n-1)}$$

erhält man für die Varianz

$$\begin{aligned}\mathrm{Var}_a[R_i] &= \frac{(N-2)(N-1)}{(n-1)(n-2)}E_a[E_z[\frac{D_{ijj'}^{(n-1)2}}{2}|R]] - \\ &\quad - \frac{(N-n)(N-1)}{(n-2)(n-1)}E_a[E_z[\frac{B_{ijj'}^{(n-1)}}{2}|R]] \\ &= E_{az}[\frac{N-1}{(n-1)(n-2)}((N-2)\frac{D_{ijj'}^{(n-1)2}}{2} - (N-n)\frac{B_{ijj'}^{(n-1)}}{2})] .\end{aligned}$$

Der Ausdruck in den eckigen Klammern stellt eine erwartungstreue Schätzung für $Var_a[R_i]$ dar. Negative Werte sind möglich.

Betrachtet man statt $Var_a[R_i]$ jetzt $Var_a[\frac{R_i}{N}]$ und ist N groß im Vergleich zu n, so ergibt sich näherungsweise:

$$Var_a[\frac{R_i}{N}] \cong E_{az}[\frac{1}{(n-1)(n-2)}(\frac{D_{ijj'}^{(n-1)2}}{2} - \frac{B_{ijj'}^{(n-1)}}{2})] .$$

Die Schätzung für $Var_a[\frac{R_i}{N}]$, welche in den eckigen Klammern steht, ist von N unabhängig.

Es soll noch eine weitere Möglichkeit der Ableitung einer erwartungstreuen Schätzung angegeben werden. Hierfür wird die erhaltene Formel für $Var_a[R_i]$ umgestellt. Man erhält

$$\frac{N-1}{n-1} E_{za}[\frac{D_{ijj'}^{(n-1)2}}{2}] = \frac{N-n}{N-2} E_a[\frac{B_{ijj'}}{2}] + \frac{n-2}{N-2} Var_a[R_i] .$$

Auf beiden Seiten wird $Var_a[R_i]$ subtrahiert:

$$\frac{N-1}{n-1} E_{za}[\frac{D_{ijj'}^{(n-1)2}}{2}] - Var_a[R_i] = \frac{N-n}{N-2} E_a[\frac{B_{ijj'}}{2}] + Var_a[R_i](\frac{n-2}{N-2} - 1) ,$$

$$= \frac{N-n}{N-2} (E_a[\frac{B_{ijj'}}{2}] - Var_a[R_i]) .$$

$$\frac{1}{N-n}(\frac{N-1}{n-1} E_{za}[\frac{D_{ijj'}^{(n-1)2}}{2}] - Var_a[R_i]) = \frac{1}{N-2}(E_a[\frac{B_{ijj'}}{2}] - Var_a[R_i]) .$$

Die rechte Seite der Gleichung ist unabhängig vom Stichprobenumfang. Betrachtet man die Gleichung für eine Zufallsstichprobe vom Umfang (n-1) und für eine weitere Zufallsstichprobe vom Umfang (m-1) ($1 < n-1$, $m-1 < N$ und $n \neq m$) und subtrahiert die beiden Gleichungen voneinander, so erhält man

$$\frac{1}{N-n}(\frac{N-1}{n-1} E_{za}[\frac{D_{ijj'}^{(n-1)2}}{2}] - Var_a[R_i]) = \frac{1}{N-m}(\frac{N-1}{m-1} E_{za}[\frac{D_{ijj'}^{(m-1)2}}{2}] - Var_a[R_i]) .$$

Durch Auflösen nach $Var_a[R_i]$ erhält man:

$$\frac{m-n}{(N-m)(N-n)} Var_a[R_i] = \frac{N-1}{(N-m)(m-1)} E_{za}[\frac{D_{ijj'}^{(m-1)2}}{2}] -$$

$$- \frac{N-1}{(N-n)(n-1)} E_{za} \left[\frac{D_{ijj'}^{(m-1)2}}{2} \right]$$

$$\mathrm{Var}_a[R_i] = E_{za} \left[\frac{N-1}{m-n} \left(\frac{N-n}{m-1} \frac{D_{ijj'}^{(m-1)2}}{2} - \frac{N-m}{n-1} \frac{D_{ijj'}^{(n-1)2}}{2} \right) \right] .$$

Der Ausdruck in den eckigen Klammern stellt eine erwartungstreue Schätzung für $\mathrm{Var}_a[R_i]$ dar. Für N groß gegen n und m ergibt sich näherungsweise

$$\mathrm{Var}_a \left[\frac{R_i}{N} \right] \cong E_{az} \left[\frac{1}{m-n} \left(\frac{1}{m-1} \frac{D_{ijj'}^{(m-1)2}}{2} - \frac{1}{n-1} \frac{D_{ijj'}^{(n-1)2}}{2} \right) \right] .$$

Hat man statt 2 jetzt k Rangordnungen von k unabhängigen homogenen Ratern, so wird für alle $\binom{k}{2}$ Paare von Rangordnungen $\frac{D_{ijj'}^{(n-1)2}}{2}$ berechnet und von diesen Werten der Mittelwert genommen:

$$\frac{1}{\binom{k}{2}} \sum_{j<j'} \frac{D_{ijj'}^{(n-1)2}}{2} .$$

Dieser Wert wird in die Schätzfunktion eingesetzt.

2.3. Schätzen der Reliabilität

2.3.1. Die N Individuen der Population werden in eine Rangreihe gebracht

Die Reliabilität als Gütemaß für den urteilenden Rater bezieht sich nicht auf ein bestimmtes festes Individuum i, sondern auf die gesamte Rangreihe.
Es seien k homogene Rater vorhanden, welche unabhängig voneinander die N Individuen der Population beurteilen.
Die Reliabilität ist definiert als

$$\rho^2[R,T] = \frac{\mathrm{Var}_g[T]}{\mathrm{Var}_g[R]} .$$

Da hier immer

$$\mathrm{Var}_g[R] = \frac{N^2-1}{12}$$

ist, hängt, wie schon bemerkt wurde, die Reliabilität nur von $Var_g[T]$, der Varianz der wahren Werte, ab.

$$Var_g[T] = Var_g[R] - Var_g[E] = Var_g[R] - E_g[E^2] =$$

$$= Var_g[R] - \frac{1}{N} \sum_i Var_a[E_i] \ .$$

Damit ist

$$\rho^2[R,T] = 1 - \frac{\frac{1}{N} \sum_i Var_a[E_i]}{Var_g[R]} = 1 - \frac{\frac{1}{N} \sum_i Var_a[R_i]}{Var_g[R]} \ .$$

Hat man eine erwartungstreue Schätzung für $Var_a[R_i]$, so auch für $\rho^2[R,T]$. Eine solche Schätzung ist (vergleiche 2.2.1.):

$$\hat{\sigma}^2(R_i) = \frac{1}{k-1} \sum_{j=1}^{k} (R_{ij} - \bar{R}_i)^2 \text{ oder}$$

$$\hat{\sigma}^2(R_i) = \frac{1}{\binom{k}{2}} \sum_{j<j'} \frac{D^2_{ijj'}}{2}$$

bei k homogenen Ratern, bzw.

$$\hat{\sigma}^2(R_i) = \frac{D^2_{ijj'}}{2}$$

bei 2 homogenen Ratern.

Damit ist z.B. (bei 2 Ratern):

$$\hat{\rho}^2[R,T] = 1 - \frac{\frac{1}{N} \sum_i \frac{D^2_{ijj'}}{2}}{Var_g[R]}$$

eine erwartungstreue Schätzung der Reliabilität.

Wird $\rho^2[R,T]$ auf diese Weise geschätzt, so ist das gleichbedeutend mit der Berechnung des Rangkorrelationskoeffizienten nach Spearman für die von den 2 Ratern genannten Rangordnungen $(j_1,\ldots,j_N)$ und $(j'_1,\ldots,j'_N)$:

$$\rho_{jj'} = 1 - \frac{6 \sum_i D^2_{ijj'}}{(N^2-1) \cdot N} \quad \text{(Kendall, 1962)}$$

$$= 1 - \frac{6 \cdot 2 \sum_i D^2_{ijj'}}{2 \cdot (N^2-1) \cdot N} = 1 - \frac{\frac{1}{N} \sum_i \frac{D^2_{ijj'}}{2}}{Var_g[R]} \quad .$$

Der Populationsparameter ρ ist

$$\rho = E_a[\rho_{jj'}] = E_a[1 - \frac{\frac{1}{N} \sum_i \frac{D^2_{ijj'}}{2}}{Var_g[R]}]$$

$$= 1 - \frac{\frac{1}{N} \sum_i E_a[\frac{D^2_{ijj'}}{2}]}{Var_g[R]} = 1 - \frac{\frac{1}{N} \sum_i Var_a[R_i]}{Var_g[R]} \quad \text{(Abschnitt 2.2.1.)}$$

$$= \rho^2[R,T] \; .$$

Dabei bedeutet E_a hier die Bildung des Erwartungswertes bezüglich der bivariaten Verteilung der beiden Rater.

$$\rho = E_a[\rho_{jj'}] = \rho^2[R,T]$$

bedeutet, daß bei Vorhandensein zweier homogener Rater die Reliabilität auch als Populationsparameter des Rangkorrelationskoeffizienten nach Spearman definiert ist. $\rho_{jj'}$ ist eine erwartungstreue Schätzung für ρ und damit für $\rho^2[R,T]$.

Der Spearman'sche Rangkorrelationskoeffizient kann auch berechnet werden aus

$$\rho_{jj'} = \frac{12}{N^2-1} \{ \frac{1}{N} \sum_{i=1}^{N} R_{ij} \cdot R_{ij'} - (\frac{N+1}{2})^2 \} \; .$$

Sind k Rater vorhanden, so kann für alle $\binom{k}{2}$ Paare von Rangordnungen der Rangkorrelationskoeffizient berechnet werden. Der Mittelwert ist dann auch eine erwartungstreue Schätzung für $\rho^2[R,T]$.

2.3.2. Die n Individuen der Stichprobe werden in eine Rangreihe gebracht

Es wird angenommen, daß eine Zufallsstichprobe vom Umfang n aus der Population vorliegt, sowie zwei homogene Rater j,j', welche die n Individuen beurteilen. Die Reliabilität bezüglich der Stichprobe (aufgefaßt als Teilpopulation) ist dann

$$\rho^2[R^{(n)},T^{(n)}] = \frac{Var_g[T^{(n)}]}{Var_g[R^{(n)}]} .$$

Analog zum Vorgehen in der Population kann $\rho^2[R^{(n)},T^{(n)}]$ geschätzt werden durch

$$\hat{\rho}^2[R^{(n)},T^{(n)}] = 1 - \frac{\frac{1}{n}\sum_i \frac{D^{(n)2}_{ijj'}}{2}}{Var_g[R^{(n)}]}$$

und es ist

$$\rho^{(n)} = E_a[\rho^{(n)}_{jj'}] = \rho^2[R^{(n)},T^{(n)}] .$$

$\rho^{(n)}_{jj'}$ ist eine erwartungstreue Schätzung für die Reliabilität $\rho^2(R^{(n)},T^{(n)})$ bezüglich der Stichprobe. In den meisten Fällen ist jedoch nicht die Reliabilität bezüglich der Stichprobe, sondern in Bezug auf die Population von Interesse, d.h. aus $\rho^{(n)}_{jj'}$ soll eine erwartungstreue Schätzung für $\rho^2[R,T]$ hergeleitet werden. $\rho^{(n)}_{jj'}$ selbst ist keine erwartungstreue Schätzung für $\rho=\rho^2[R,T]$, wie man den folgenden Ausführungen entnehmen kann.

Eine Möglichkeit, eine erwartungstreue Schätzung abzuleiten, bietet die Gleichung von Durbin, Stuart und Daniels (Kendall, 1962, S. 93) für zwei feste Rangreihen:

$$E_z[\rho^{(n)}_{jj'}|R] = \frac{1}{(n+1)(N-2)}\{3(N-n)\tau_{jj'}+(n-2)(N+1)\rho_{jj'}\} .$$

Die Gleichung stellt eine Beziehung her zwischen dem Rangkorrelationskoeffizienten nach Spearman bezüglich der Stich-

probe, $\rho_{jj'}^{(n)}$, und den Koeffizienten $\rho_{jj'}$ und $\tau_{jj'}$ (Korrelation nach Kendall) in Bezug auf die Population. Man kann sich hierfür vorstellen, daß die beiden homogenen Rater j, j' je eine Rangreihe der N Individuen der Population nannten. $\tau_{jj'}$ und $\rho_{jj'}$ könnten berechnet werden. Zieht man dann aus der Population eine Zufallsstichprobe vom Umfang n, so können $\rho_{jj'}^{(n)}$ und $\tau_{jj'}^{(n)}$ berechnet werden. Es gilt (Kendall, 1962, S. 81):

$$E_z[\tau_{jj'}^{(n)}|R] = \tau_{jj'} .$$

Die Gleichung

$$E_z[\rho_{jj'}^{(n)}|R] = \frac{1}{(n+1)(N-2)}\{3(N-n)\tau_{jj'}+(n-2)(N+1)\rho_{jj'}\}$$

wird nach $\rho_{jj'}$ aufgelöst:

$$\rho_{jj'} = \frac{1}{(n-2)(N+1)}\{(n+1)(N-2)E_z[\rho_{jj'}^{(n)}|R]-3(N-n)\tau_{jj'}\} .$$

Bilden des Erwartungswertes bezüglich der Rangordnungen ergibt (mit $E_a[\rho_{jj'}]=\rho=\rho^2[R,T]$):

$$\rho = \frac{1}{(n-2)(N+1)}\{(n+1)(N-2)E_aE_z[\rho_{jj'}^{(n)}|R]-3(N-n)E_a[\tau_{jj'}]\} ,$$

$$\rho^2[R,T] = E_aE_z\left[\frac{1}{(n-2)(N+1)}\{(n+1)(N-2)\rho_{jj'}^{(n)}-3(N-n)\tau_{jj'}^{(n)}\}|R\right]$$

$$= E_{az}\left[\frac{1}{(n-2)(N+1)}\{(n+1)(N-2)\rho_{jj'}^{(n)}-3(N-n)\tau_{jj'}^{(n)}\}\right] .$$

Der Ausdruck in den eckigen Klammern stellt eine erwartungstreue Schätzung für $\rho^2[R,T]$ dar.

Ist N sehr groß im Vergleich zu n, so erhält man näherungsweise die Schätzung:

$$\rho^2[R,T] \cong E_{az}\left[\frac{1}{n-2}((n+1)\rho_{jj'}^{(n)}-3\tau_{jj'}^{(n)})\right] .$$

Die Schätzfunktion ist unabhängig von N.

Will man $\rho^2[R,T]$ nur mit $\tau_{jj}^{(n)}$, oder mit $\rho_{jj}^{(n)}$, schätzen, so ist dem Rangkorrelationskoeffizienten $\rho_{jj}^{(n)}$, nach Spearman der Vorzug zu geben, denn nach Auflösung der obenstehenden Gleichung für $\rho^2[R,T]$ nach $E_{az}[\tau_{jj}^{(n)},]$ und Subtraktion von $\rho^2[R,T]$ auf beiden Seiten der Gleichung ergibt sich

$$E_{az}[\tau_{jj}^{(n)},]-\rho^2[R,T] = \frac{(N-2)(n+1)}{3(N-n)} \; (E_{az}[\rho_{jj}^{(n)},]-\rho^2[R,T]) \; .$$

Da $\frac{(N-2)(n+1)}{3(N-n)} > 1$ ist für $n > 2$, ist der Betrag der Differenz auf der rechten Seite (innerhalb der Klammern) kleiner als der Betrag der Differenz auf der linken Seite und damit ist $\rho_{jj}^{(n)}$, als Schätzer zu bevorzugen.

Eine andere Möglichkeit, eine erwartungstreue Schätzung zu erhalten, wird im folgenden angegeben.

Aus der Beziehung für die Varianz der Rangzahlen eines Individuums (Abschnitt 2.2.2.)

$$\frac{1}{N-n}\{\frac{N-1}{n-1} E_z[Var_a[R_i^{(n-1)}]]-Var_a[R_i]\} =$$

$$= \frac{1}{N-m}\{\frac{N-1}{m-1} E_z[Var_a[R_i^{(m-1)}]]-Var_a[R_i]\}$$

kann eine ähnliche für $\rho^2[R,T]$ und $\rho_{jj}^{(n)}$, abgeleitet werden.

Zu diesem Zweck wird nicht mehr ein festes Individuum i betrachtet, sondern ein mit der Wahrscheinlichkeit 1/N aus der Population ausgewähltes. Formal bedeutet dieses, daß der Erwartungswert bezüglich der Zufallsauswahl der Individuen gebildet wird:

$$\frac{1}{N-n}\{\frac{N-1}{n-1} E_z E_i[Var_a[R_i^{(n-1)}]]-E_i[Var_a[R_i]]\} =$$

$$= \frac{1}{N-m}\{\frac{N-1}{n-1} E_z E_i[Var_a[R_i^{(n-1)}]]-E_i[\,Var_a[R_i]]\} \; .$$

E_i und E_z sind miteinander vertauschbar, da E_i hier die

Summierung von i=1 bis i=N und Division durch N bedeutet, und eine endliche Summe mit dem Erwartungswert E_z vertauscht werden kann. Man kann sich vorstellen, daß zuerst das Individuum mit der Wahrscheinlichkeit 1/N aus der Population ausgewählt und dann die Zufallsstichprobe vom Umfang (n-1) bzw. (m-1) gezogen wird. Das ergibt eine Zufallsstichprobe vom Umfang n bzw. m.

Mit

$$E_i[Var_a[R_i]] = E_i[E_a[E^2|I]] = E_g[E^2] = Var_g[E] \quad \text{und}$$

$$E_i[Var_a[R_i^{(n)}]] = E_i[E_a[E^{(n)2}|I]] = E_g[E^{(n)2}] = Var_g[E^{(n)}]$$

ist

$$\frac{1}{N-n}\{\frac{N-1}{n-1} E_z[Var_g[E^{(n)}]]-Var_g[E]\} =$$

$$= \frac{1}{N-m}\{\frac{N-1}{m-1} E_z[Var_g[E^{(m)}]]-Var_g[E]\} .$$

Mit

$$Var_g[E] = Var_g[R](1-\rho^2[R,T]),$$

$$Var_g[E^{(n)}] = Var_g[R^{(n)}](1-\rho^2[R^{(n)},T^{(n)}])$$

ist

$$\frac{1}{N-n}\{\frac{N-1}{n-1} Var_g[R^{(n)}]E_z[1-\rho^2[R^{(n)},T^{(n)}]]-$$

$$-Var_g[R](1-\rho^2[R,T])\} =$$

$$= \frac{1}{N-m}\{\frac{N-1}{m-1} Var_g[R^{(m)}]E_z[1-\rho^2[R^{(m)},T^{(m)}]]-$$

$$-Var_g[R](1-\rho^2[R,T])\} .$$

Mit $Var_g[R] = \frac{N^2-1}{12}$ und $Var_g[R^{(n)}] = \frac{n^2-1}{12}$ ist

$$\frac{1}{N-n}\left\{\frac{N-1}{n-1}\cdot\frac{n^2-1}{12}\,E_z[1-\rho^2[R^{(n)},T^{(n)}]]-\frac{N^2-1}{12}\,(1-\rho^2[R,T])\right\}=$$

$$=\frac{1}{N-m}\left\{\frac{N-1}{m-1}\,\frac{m^2-1}{12}\,E_z[1-\rho^2[R^{(m)},T^{(m)}]]-\frac{N^2-1}{12}\,(1-\rho^2[R,T])\right\},$$

$$\frac{1}{N-n}\left\{(n+1)E_z[1-\rho^2[R^{(n)},T^{(n)}]]-(N+1)(1-\rho^2[R,T])\right\}=$$

$$=\frac{1}{N-m}\left\{(m+1)E_z[1-\rho^2[R^{(m)},T^{(m)}]]-(N+1)(1-\rho^2[R,T])\right\},$$

$$\frac{1}{N-n}\left\{n-N+(N+1)\rho^2[R,T]-(1+n)E_z[\rho^2[R^{(n)},T^{(n)}]]\right\}=$$

$$=\frac{1}{N-m}\left\{m-N+(N+1)\rho^2[R,T]-(1+m)E_z[\rho^2[R^{(m)},T^{(m)}]]\right\},$$

$$\frac{(N-m)(n-N)-(m-N)(N-n)}{(N-n)(N-m)}+\left(\frac{N+1}{N-n}-\frac{N+1}{N-m}\right)\rho^2[R,T]=$$

$$=\frac{1+n}{N-n}\,E_z[\rho^2[R^{(n)},T^{(n)}]]-\frac{1+m}{N-m}\,E_z[\rho^2[R^{(m)},T^{(m)}]]$$

$$\rho^2[R,T]=\frac{(N-n)(N-m)}{(N+1)(n-m)}\left\{\frac{1+n}{N-n}\,E_z[\rho^2[R^{(n)},T^{(n)}]]-\right.$$

$$\left.-\frac{1+m}{N-m}\,E_z[\rho^2[R^{(m)},T^{(m)}]]\right\}$$

$$=\frac{1}{(N+1)(n-m)}\left\{(N-m)(n+1)E_z[\rho^2[R^{(n)},T^{(n)}]]-\right.$$

$$\left.-(N-n)(m+1)E_z[\rho^2[R^{(m)},T^{(m)}]]\right\},$$

$$\rho^2[R,T]=E_z\left[\frac{1}{(N+1)(n-m)}\left((N-m)(n+1)\rho^2[R^{(n)},T^{(n)}]-\right.\right.$$

$$\left.\left.-(N-n)(m+1)\rho^2[R^{(m)},T^{(m)}]\right)\right].$$

Der Ausdruck in den eckigen Klammern stellt ebenfalls eine erwartungstreue Schätzung für $\rho^2[R,T]$ dar.

Für N groß gegen n und m erhält man näherungsweise

$$\rho^2[R,T]\cong E_z\left[\frac{1}{n-m}\left\{(n+1)\rho^2[R^{(n)},T^{(n)}]-(m+1)\rho^2[R^{(m)},T^{(m)}]\right\}\right].$$

2.4. Schätzen der Korrelation zweier ordinaler Tests

2.4.1. Die N Individuen der Population werden in eine Rangreihe gebracht

Die Korrelation zweier ordinaler Tests wird z.B. zum Schätzen der Validität oder Reliabilität benötigt.
Es wird angenommen, daß 2 homogene Rater j, j' die Rangreihen $(j_1,\ldots,j_N)$ und $(j'_1,\ldots,j'_N)$ nennen.

Eine erwartungstreue Schätzung für die Korrelation in der Population $\rho = E_a[\rho_{jj'}]$ ist

$$\rho_{jj'} = \frac{12}{N^2-1}\left\{\frac{1}{N}\sum_{i=1}^{N} R_{ij}\cdot R_{ij'} - \left(\frac{N+1}{2}\right)^2\right\},$$

welches die Berechnung der Rangkorrelationskoeffizienten nach Spearman bedeutet (siehe auch Abschnitt 2.3.1.).

2.4.2. Die n Individuen der Stichprobe werden in eine Rangreihe gebracht

Es wird angenommen, daß 2 homogene Rater j, j' sowie eine Zufallsstichprobe vom Umfang n aus der Population vorliegen. Aus der Korrelation bezüglich der Stichprobe soll die Korrelation $\rho[R,R']$ bezüglich der Population geschätzt werden.

Das Problem ist analog zu demjenigen des Abschnitts 2.3.2. Die erwartungstreuen Schätzungen werden analog abgeleitet und daher hier nur angegeben:

Mit der Gleichung von Durbin, Stuart und Daniels ergibt sich

$$\rho[R,R'] = E_{az}\left[\frac{1}{(n-2)(N+1)}\left\{(n+1)(N-2)\rho^{(n)}_{jj'} - 3(N-n)\tau^{(n)}_{jj'}\right\}\right].$$

Der Ausdruck in den eckigen Klammern stellt eine erwartungstreue Schätzung für $\rho[R,R']$ dar.

Unter Verwendung von 2 Zufallsstichproben erhält man

$$\rho[R,R'] = E_z \left[\frac{1}{(N+1)(n-m)} \left\{ (N-m)(n+1)\rho^{(n)} - (N-n)\cdot(m+1)\rho^{(m)} \right\} \right] ,$$

wobei der Ausdruck in den eckigen Klammern eine erwartungstreue Schätzung darstellt.

Auch hier erhält man für N sehr groß im Vergleich zu n bzw. n und m

$$\rho[R,R'] \cong E_{az} \left[\frac{1}{n-2} \left((n+1)\rho_{jj'}^{(n)} - 3\tau_{jj'}^{(n)} \right) \right]$$

bzw.

$$\rho[R,R'] \cong E_z \left[\frac{1}{n-m} \left\{ (n+1)\rho^{(n)} - (m+1)\rho^{(m)} \right\} \right] .$$

2.5. Zusammenfassung

Für die wichtigsten Konzepte der KTT wurde gezeigt, daß sie auf Rangzahlen übertragen werden können. Dieses liegt an der Definition des wahren Wertes als Erwartungswert und der sich damit ergebenden Aufspaltung der Rangzahl in einen wahren Wert und in einen Meßfehler. Das Nachprüfen der Konzepte und vor allem die Schätzung erwiesen sich als z.T. recht aufwendig.

Statt der Rangzahlen R_i der Individuen i können auch die relativen Rangzahlen R_i/N, i=1,...,N, betrachtet werden. Die Betrachtung der relativen Ränge bedeutet eine Normierung. Dieses Vorgehen ist dann zu empfehlen, wenn der Populationsumfang N groß gegen n ist. Dann verschwindet näherungsweise die Abhängigkeit der Schätzfunktionen von N.

In praktischen Situationen wird häufig der Umfang N der Population unbekannt sein. Dann ist weder das Ziehen einer Zufallsstichprobe noch eine Schätzung möglich. Die beste Lösung wäre dann, die Stichprobe als "repräsentative Stichprobe" aufzufassen und diese wiederum der unbekannten Population gleichzusetzen. Wie der vorangegangene Abschnitt gezeigt hat, sind bei bekannter Population die Schätzungen einfach.

Für die Ableitung der Formeln im vorangegangenen Abschnitt wurde vorausgesetzt, daß keine Bindungen vorliegen. Sind Bindungen vorhanden, so kann für die Varianz $Var_g[R]$ kein geschlossener Ausdruck mehr angegeben werden, da eine Abhängigkeit von der jeweils vorliegenden Bindungskonstellation besteht. Das Ergebnis $Var_g[R] = (N^2-1)/12$ beruhte auf der Forderung, daß keine Bindungen vorhanden sind. Betrachtet man die Formeln zur Schätzung des wahren Wertes, der Varianz und der Reliabilität, so fällt auf, daß diese Varianz nur bei der Schätzung der Reliabilität in der Population benutzt wurde. Statt dessen kann eine Schätzung der Reliabilität mit Hilfe von Paralleltests erfolgen. Die übrigen Schätzformeln, z.B. zur Schätzung des wahren Wertes t_i, der

Varianz $Var_a[R_i]$ und der Korrelation zweier Tests können auch benutzt werden, wenn Bindungen vorhanden sind und gleichen Werten Mittelränge zugeordnet werden. Werden die Rangzahlen den gleichen Werten zufällig zugewiesen, so gilt auch weiterhin $Var_g[R] = (N^2-1)/12$.

Ein ordinaler Test wurde hier durch einen Rater repräsentiert. Das ordinale Testmodell kann als einfaches Urteilsmodell für einen Rater aufgefaßt werden.
Im folgenden werden einige Vorschläge für seine Anwendung gemacht.

3. Anwendung des ordinalen Testmodells

3.1. Skalieren von Items

Das ordinale Testmodell soll zur Konstruktion einer Rangskala verwendet werden. Beispielsweise möge es sich um eine Einstellungsskala handeln. Die zu beurteilenden Objekte sind N Items, welche die Ausprägungen des betrachteten Merkmals repräsentieren. Die Items werden bezüglich ihrer Merkmalsausprägung in eine Rangordnung gebracht. Damit entsprechen die N Items der Population der N Individuen im ordinalen Testmodell.
Wenn man die N Items als Population statt als Zufallsstichprobe auffaßt, hat das den Vorteil, daß die Schätzungen einfacher sind. Man kann sich vorstellen, daß es sich um eine repräsentative Stichprobe aus einer Population von Items handelt. Die Stichprobe wird als "Eichpopulation" aufgefaßt und diese gleich der Population gesetzt. Aus dieser Population von N Items soll eine kleinere Anzahl nach gewissen Gütekriterien ausgelesen werden. Der Vorgang der Auswahl geeigneter Items ist als Itemanalyse bekannt.
Es sei darauf hingewiesen, daß es sich bei der zu erstellende Skala um eine abgeleitete Skala handelt. Die repräsentierende Relation $t_i = E[R_i]$, welche dem Item i, i=1,...,N, einen wahre Skalenwert zuordnet, ist empirisch nicht überprüfbar. Die fun mentale Messung nehmen die Rater bei der Beurteilung vor.

3.1.1. Vorgehen

Analog zur klassischen eindimensionalen Skalierung und zur KT ist es auch hier die erste Aufgabe, Items zu formulieren. Die Items sollen unterschiedliche Ausprägungsgrade des betrachtet Merkmals repräsentieren. Die Formulierung der Items und die A wortvorgabe für eine spätere Beantwortung durch Probanden sin abhängig von der Art des betrachteten Merkmals. Für eine Eins lungsskala werden üblicherweise die Items als Behauptungen fo liert. Die endgültige Skala soll aus einer gewissen Anzahl vo Items bestehen, die bezüglich ihrer Ausprägung auf dem betrac ten Merkmal eine Rangordnung bilden. Der Skalenwert des Items eine Rangzahl sein.

Für die Konstruktion der Skala bringen g (g>1), homogene Rater unabhängig voneinander die N Items, welche Ausprägungen des betrachteten Merkmals repräsentieren, in g Rangreihen. Bindungen seien nicht vorhanden. Unter homogenen Ratern werden solche verstanden, welche dem Item i, i=1,...,N, gleiche wahre Werte und gleiche Varianzen der Rangzahlen zuweisen.

Aus den g Rangreihen wird pro Item eine Häufigkeitsverteilung für die Rangzahlen des Items abgeleitet, welche durch den Mittelwert und die empirische Varianz der Rangzahlen gekennzeichnet wird. Der Rangmittelwert

$$\bar{R}_i = \frac{1}{g} \sum_{i=1}^{g} R_{ij}$$

des Items i ist eine erwartungstreue Schätzung des wahren Wertes t_i des Items i und

$$\hat{\sigma}^2(R_i) = \frac{1}{g-1} \sum_{j=1}^{g} (R_{ij}-\bar{R}_i)^2$$

ist eine erwartungstreue Schätzung der Varianz der Rangzahlen $Var_a[R_i]$ des Items. Der wahre Wert und die Varianz der Rangzahlen eines Items sind Itemkennzahlen. Der wahre Skalenwert des Items i wird als sein wahrer Wert t_i definiert. Damit wird der wahre Skalenwert als der Erwartungswert der Wahrscheinlichkeitsverteilung für das Item definiert. Diese Wahrscheinlichkeitsverteilung wird aus der Urteilsverteilung des Raters für die N! möglichen Rangreihen abgeleitet. Aus den Häufigkeitsverteilungen hat man somit pro Item Schätzungen des wahren Skalenwertes und der Varianz der Rangzahlen erhalten.

3.1.2. Itemanalyse

Für eine reliable Skala werden Items benötigt, deren wahre Werte eine möglichst große Varianz aufweisen. Da die wahren Werte unbekannt sind, werden für die Itemauswahl die geschätzten Werte verwendet.
Die Auswahl erfolgt so, daß die auszuwählenden Items sich

möglichst "gleichmäßig" über die Rangskala verteilen, d.h. es sollen möglichst alle Rangbereiche vertreten sein. Für diesen ersten Schritt der Itemanalyse werden pro Item die Schätzung des wahren Skalenwertes und die empirische Varianz der Rangzahlen berücksichtigt. Von zwei Items mit gleichen oder nahezu gleichen geschätzten Skalenwerten wird dasjenige mit der geringeren Varianz ausgewählt. Die so ausgewählten N' Items bilden eine vorläufige Skala.

In einem zweiten Schritt der Itemanalyse wird die erstellte Skalenvorform mehreren Probanden zur Beantwortung vorgelegt. Die Items können mit "stimmt" oder "stimmt nicht" beantwortet werden. Als weiteres Kriterium soll jetzt die Trennschärfe der Items berücksichtigt werden. Im Rahmen des ordinalen Testmodells soll unter der "Trennschärfe des Items i" der Item-Test-Zusammenhang verstanden werden, wobei der "Zusammenhang" noch zu definieren ist. Um die Trennschärfe der Items schätzen zu können, müssen somit Probanden die vorläufige Skala, die aus den N' Items besteht, beantworten, und zwar nach ihrer eigenen Einstellung. Der Score des Probanden wird als der Median der Schätzungen der wahren Skalenwerte der positiv beantworteten Items gewählt. Während für den ersten Schritt der Itemanalyse g homogene, objektiv und unabhängig voneinander urteilende Rater erforderlich waren, sind es hier Probanden, die nach ihrer eigenen Einstellung antworten.

Soll die Skala Items aus allen Skalenbereichen enthalten, so kann die Trennschärfe durch die Rangkorrelation nach Spearman zwischen den Antworten der Probanden auf das Item und den Scores der Probanden geschätzt werden. Bei der Itemanalyse werden Items mit hohen Korrelationen aus allen Skalenbereichen ausgewählt. Sollen die Items insbesondere zwischen extremen Gruppen differenzieren, so wird ein Extremgruppenvergleich durchgeführt. Hierfür wird die Häufigkeitsverteilung der Scores der Probanden erstellt. Um Extremgruppen zu erhalten, werden z.B. aus dieser Verteilung diejenigen 25% der Probanden bestimmt, welche die höchsten Scores haben und ebenso diejenigen 25% mit den niedrigsten Scores. Gibt jeder Proband die Antwort "stimmt" oder "stimmt nicht" ab, so kann pro Item eine Vierfeldertafel erstellt werden:

Tab. 1.1.: Zusammenhang von Antworten der Probanden und Gruppenzugehörigkeit

Gruppe	Antwort stimmt	stimmt nicht	
obere 25%	n_{11}	n_{12}	$n_{1.}$
untere 25%	n_{21}	n_{22}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	$n_{..}$

Der Zusammenhang von Itemantwort und Gruppenzugehörigkeit wird durch Φ oder Φ^2 angegeben, mit

$$\Phi^2 = \frac{(n_{11}n_{22}-n_{12}n_{21})^2}{n_{1.}n_{.1}n_{2.}n_{.2}} \quad .$$

In diesem Fall wird die Trennschärfe durch ϕ^2 $(0 \leq \phi^2 \leq 1)$ geschätzt. Items mit großen Werten für ϕ^2 werden für die Skala ausgewählt. Bei einem Extremgruppenvergleich werden in der Regel Items aus dem mittleren Skalenbereich eliminiert, da sie nicht genügend zwischen den Extremgruppen differenzieren, was zu einem niedrigen ϕ^2 führt. Den schließlich für die endgültige Skala ausgelesenen N'' Items werden nach der Höhe der Schätzungen der wahren Skalenwerte Rangzahlen zugeordnet. Der Skalenwert des Items i wird als seine Rangzahl R_i definiert. Damit liegt eine Rangskala vor.

3.1.3. Schätzen der Reliabilität und Validität der Rater

Die Reliabilität kann geschätzt werden, indem die ausgewählten N'' Items z.B. zwei homogenen Ratern, welche unabhängig voneinander urteilen, vorgegeben werden. Diese sollen jeweils die Items in eine Rangreihe bringen. Die Reliabilität kann nach Abschnitt I.2.3.1. erwartungstreu geschätzt werden mit dem Rangkorrelationskoeffizienten nach Spearman

$$\hat{\rho}^2[R,T] = 1 - \frac{6\sum_i D_{ijk}^2}{(N^2-1)N} \quad .$$

Um die Validität schätzen zu können, muß ein weiterer Rater die

Items bezüglich einer Kriteriumsvariablen beurteilen und in eine Rangreihe bringen. Jedes Item hat dann zwei Rangzahlen, eine für seine Ausprägung bezüglich des betrachteten Merkmals, die andere für seine Ausprägung auf der Kriteriumsvariablen. Die Validität kann ebenfalls erwartungstreu mit Hilfe des Rangkorrelationskoeffizienten nach Spearman geschätzt werden (Abschnitt I.2.4.1.).

3.1.4. Einordnen von Individuen auf dieser Skala

Um einen Score für ein Individuum zu erhalten, muß dieses die N'' Items, welche die Skala bilden, mit "stimmt" bzw. "stimmt nicht" beantworten. Meistens wird es so sein, daß das Individuum mehrere Items mit "stimmt" ankreuzen wird, obwohl die Items unterschiedliche Ausprägungsgrade repräsentieren. Der Score des Individuums kann dann als Mittelwert der Rangzahlen oder auch als Median der Rangzahlen der mit "stimmt" beantworteten Items angegeben werden.

Die Stellung des Individuums zur Population kann aus seinem Score nicht abgelesen werden. Liegt z.B. eine experimentelle Situation vor und soll ein Zweigruppenvergleich durchgeführt werden, so ist der Bezug zur Population unwichtig. Anders ist es jedoch in der diagnostischen Situation. Hier muß, analog zur KTT, eine Eichpopulation erhoben werden, mit deren Hilfe man feststellen kann, wie häufig jeder Score in der Population vorkommt. Ebenfalls analog zur KTT kann für jedes Individuum sein Prozentrang berechnet werden. Wenn also 10% der Eichpopulation einen Score erhalten haben, der kleiner oder gleich dem des Individuums i ist, so erhält dieses den Prozentrang 0,1. Jedem Individuum werden zwei Zahlen zugeordnet, der erhaltene Score und der Prozentrang.

3.1.5. Testprofile

Soll ein Testprofil erstellt werden, wobei jede Skala unterschiedliche Aspekte des betrachteten Konstruktes erfaßt, so wird jede dieser Skalen nach den angegebenen Methoden konstruiert. Das Individuum erhält statt eines Scores einen Vektor, das Testprofil, zugeordnet.

3.1.6. Konstruktion einer Leistungsskala

Die Skalenkonstruktion mit Hilfe des ordinalen Testmodells ist nicht auf Einstellungsskalen beschränkt. Es ist auch möglich, auf diese Art z.B. eine Leistungsskala zu erstellen. Die Voraussetzung hierfür ist, daß sich die Items nach einem Kriterium in eine Rangreihe bringen lassen. Ein solches Kriterium kann die Schwierigkeit der zu lösenden Aufgabe sein. Die "Schwierigkeit" kann einmal normativ aufgefaßt werden, z.B. bestimmt ein Mathematiklehrer als Rater wie schwer die Aufgabe für die Klassenarbeit ist. Eine andere Möglichkeit ist es, die Schwierigkeit empirisch zu bestimmen, analog zur KTT. Eine Schätzung der Aufgabenleichtigkeit ist dann die relative Häufigkeit der Individuen in der Population, welche die betreffende Aufgabe gelöst haben. Es ist Aufgabenschwierigkeit = 1-Leichtigkeit der Aufgabe.

Die Aufgabenerstellung sowie die Art der Aufgabenbewertung bedürfen einer eingehenden Diskussion. Sollen die Aufgaben direkt in eine Rangreihe gebracht werden oder erhalten sie zunächst Punkte, die dann auf Rangzahlen abgebildet werden? Welche Leistungsbereiche sollen erfaßt werden, wie gut sollen die Skalen differenzieren?
Alle diese Fragen können nur in Zusammenhang mit der jeweils vorliegenden Fragestellung gelöst werden.

Möglichkeiten für die Erstellung einer Rangreihe der Aufgaben sind:

1. g unabhängig voneinander urteilende, homogene Rater (Experten) beurteilen die Items daraufhin, wie schwer sie sind und bringen sie nach diesem Kriterium in eine Rangreihe. Die Aufgaben sollten möglichst von anderen Experten erstellt worden sein.
   Die g Rater liefern g Rangreihen der Aufgaben und man kann, wie bei den Einstellungsskalen, die wahren Werte und Varianzen der Rangzahlen der Aufgaben schätzen. Ebenso kann die Reliabilität geschätzt und die Itemanalyse durchgeführt werden.
   Die Rangreihe der Aufgaben entsteht hier nur durch das Urteil der Experten, also normativ.

2. Die Schwierigkeit der Aufgaben wird empirisch bestimmt. Die Aufgaben werden nach diesen empirischen Werten in eine Rangreihe gebracht. Eine Itemanalyse wie unter 1. ist nicht möglich da nur eine Rangreihe entsteht.

Das Individuum bekommt die Aufgaben, deren Skalenwerte Rangzahlen sind, zum Lösen vorgelegt. Als Score kann die Rangzahl der zuletzt gelösten Aufgabe gewählt werden. In der praktischen Anwendung kann es vorkommen, daß das Individuum eine Aufgabe mit der Rangzahl R_l nicht löst und eine Aufgabe mit der Rangzahl R_k und $R_l < R_k$ löst. Es kann trotzdem als Score die Rangzahl der zuletzt gelösten Aufgabe erhalten. Die zwischendurch nicht gelösten Aufgaben können extra berücksichtigt werden. Sie deuten z.B. auf mangelnde Konzentration während des Lösens der Aufgaben hin.

3.1.7. Praktikabilität der beschriebenen Skalenkonstruktion

Bei der Anwendung des ordinalen Testmodells werden die Items pro Rater in eine Rangreihe gebracht. Die Rangreihe soll möglichst keine Bindungen enthalten. Bei einer großen Anzahl von Items ist die Erstellung einer Rangreihe eine schwierige Aufgabe für die Rater, da sehr viele Vergleiche durchgeführt werden müssen. Dieses wird der Fall sein, wenn die zu erstellende Skala sehr umfangreich sein soll. Wenn die endgültige Skala etwa 30 Items enthalten soll so ist es durchaus üblich, von N=100 Items auszugehen. Hier sollte den Ratern zu einem sukzessiven Vorgehen bei der Beurteilung geraten werden. Dieses kann so aussehen, daß die Rater zunächst die Items grob in Kategorien einteilen und diese anschließend immer mehr verfeinern. Das sukzessive Vorgehen stellt eine Urteilserleichterung dar. Die Rater haben weniger Vergleiche durchzuführen, wenn eine Skala mit nur wenigen Items, z.B. also 5-10 Items, erstellt werden soll. Die Anwendung des ordinalen Testmodells kann in diesem Falle empfohlen werden.

3.2. Skalieren von Individuen

Während im vorigen Abschnitt das ordinale Testmodell zur Skalierung von Items angewandt wurde, bezieht sich die Anwendung jetzt auf das Erstellen von Rangreihen von Individuen. Beim Skalieren von Items repräsentierte ein Rater einen "ordinalen

Test" im Sinne von Schulman und Haden (1975). Er urteilte nach einer Wahrscheinlichkeitsverteilung für die Rangreihen. Alle Konzepte (z.B. Reliabilität, Validität) galten für die Rater.

Beim Skalieren von Individuen können die Rangreihen wie folgt zustandekommen:

1. g unabhängig voneinander urteilende, homogene Rater beurteilen die Individuen bezüglich der Ausprägungen des betrachteten Merkmals. Jeder Rater erstellt eine Rangreihe der Individuen. Bei dieser Anwendung wird keine Skala konstruiert.
2. Die Individuen beantworten Items. Aufgrund der erhaltenen Scores werden die Items in eine Rangreihe gebracht.

In beiden Fällen bilden die Individuen die Population des ordinalen Testmodells (im letzten Abschnitt waren es die Items). Im folgenden sollen die beiden Möglichkeiten 1. und 2. diskutiert werden.

3.2.1. Rater beurteilen Individuen

Der Fall, daß g Rater eine Population von N Individuen beurteilen, wird in der alltäglichen diagnostischen Situation selten vorkommen, da hier meistens nur ein Individuum zu beurteilen ist. Die Anwendung ist auf Situationen beschränkt, in denen eine Bezugsgruppe vorhanden ist. Ein Beispiel hierfür ist die Eignungsbeurteilung von Individuen. Auch bei sportlichen Wettkämpfen ist das ordinale Testmodell anwendbar. (z.B. Kunstturnen, Eiskunstlauf, Paartanz). Die Wettbewerbsteilnehmer erhalten in der Regel nach ihrer Leistung von mehreren Ratern Punkte zugeordnet. Aus der Gesamtpunktzahl wird dann eine Rangreihe gebildet. Ebenso könnten die Rater, statt Punkte zuzuordnen, jeweils eine Rangreihe nennen. Auch auf Leistungsbeurteilungen in der Schule ist das Modell anwendbar. Hier hängt es von der jeweiligen Fragestellung ab, wie die Population definiert wird (z.B. die Schüler der 8. Klasse einer bestimmten Schule oder einer Stadt oder eines Bundeslandes). Ein gutes Beispiel ist der Sportunterricht, wo die Schüler in jeder Disziplin

in eine Rangreihe gebracht werden und entweder mit dem Konzept des zusammengesetzten Tests Gesamtscores erhalten oder einen Vektor im Sinne eines Testprofils zugeordnet bekommen.

Mit Hilfe g homogener Rater, welche unabhängig voneinander die N Individuen in Rangreihen bringen, können für jedes Individuum der wahre Wert und seine Varianz geschätzt werden. Werden die N Individuen nicht als Population aufgefaßt, sondern nur als Stichprobe, so kann eine "Hochrechnung" auf eine Population bekannter Größe geschehen (Schätzmethoden des Abschnittes I.2.). Auch die Reliabilität und die Validität der Rater kann geschätzt werden.

Die Anwendung des ordinalen Testmodells ist nicht auf Leistungssituationen beschränkt. Auch in der Klinik kommt es z.B. vor, daß Rater Patienten beurteilen. Eine spezielle Anwendung können Einzelfallstudien sein. Die Fragestellung ist dann so, daß Zustände des betreffenden Individuums über die Zeit hinweg daraufhin beurteilt werden, ob z.B. eine Verbesserung oder eine Verschlechterung oder allgemein ein bestimmter Verlauf vorliegt. Die Anwendung des ordinalen Testmodells auf den Einzelfall bedeutet, daß statt N Individuen jetzt N Zustände eines Individuums betrachtet werden. g homogene Rater bringen also statt N Individuen unabhängig voneinander N Zustände eines Individuums in Rangreihen. Dieses kann geschehen, indem zu jedem Zeitpunkt ein Videogerät eingesetzt wird. Den g Ratern werden in zufälliger Reihenfolge die N Zustände des Individuums gezeigt. Die Rater ordnen den Zuständen Rangzahlen zu (audiovisuelle Technik, siehe z.B. Fischer-Cornelssen und Abt, 1980). Pro Zeitpunkt können für das Individuum wahre Werte und Varianzen geschätzt werden.

3.2.2. Individuen beantworten Items

Es soll eine weitere Möglichkeit zur Erstellung einer Skala erwähnt werden. Im Unterschied zum Vorgehen unter 3.1. beurteilen jetzt nicht g Rater objektiv eine Population von N Items, sondern N Individuen beantworten k Items. Es handelt sich hier um Selbstbeurteilungen von Individuen. Die N Individuen repräsentieren die Population des ordinalen Testmodells. Die Stelle der Rater nehmen jetzt die Items ein. An

die Stelle der Urteilsverteilung des Raters tritt die Selbstbeurteilungsverteilung des Individuums. Pro Item werden die Individuen bezüglich ihrer Scores, die sie für die Beantwortung des Items erhalten, in eine Rangreihe gebracht. Bei k Items entstehen k Rangreihen. Es muß sich um homogene Items handeln, die unabhängig voneinander zu beantworten sind. Homogen bedeutet, daß die Items einem festen Individuum dieselben wahren Werte und Varianzen zuordnen. Der Unterschied der beiden Vorgehensweisen ist ähnlich zu sehen wie derjenige zwischen der Methode der gleicherscheinenden Intervalle nach Thurstone und der Likert-Skalierung in der eindimensionalen Skalierung. Bei der Methode der gleicherscheinenden Intervalle beurteilen Rater objektiv die Items nach ihrer Ausprägung auf einer K-Punkte-Skala. Für die Items werden Skalenwerte berechnet. Bei der Likert-Skalierung beantworten Individuen die Items nach ihrer eigenen Einstellung. Die Items erhalten keine Skalenwerte.
Die Methoden der Abschnitte I.1. und I.2. sind direkt anwendbar. Ordinale Tests werden jetzt durch Items repräsentiert.

Für die Schätzmethoden des Abschnitts I.2. sind mindestens zwei Rangreihen erforderlich. Damit müssen zwei Items vorliegen, die dasselbe Merkmal messen.

3.2.2.1. Vorgehen

Es wird angenommen, daß k Items, welche das betrachtete Merkmal erfassen sollen, vorliegen. N Individuen beantworten die Items. Pro Item werden die Individuen nach den erhaltenen Scores jeweils in eine Rangreihe gebracht, so daß k Rangreihen entstehen. Bei gleichen Scores werden Mittelränge gebildet. Jedes Indivduum i erhält k Rangzahlen. Durch Addition entsteht ein Gesamtscore.

3.2.2.2. Itemanalyse

Für das ordinale Testmodell können nur solche Items verwendet werden, mit deren Hilfe die N Individuen in eine Rangreihe gebracht werden können, ohne daß sich zu viele Bindungen ergeben. Damit kann ein einziges

dichotom zu beantwortendes Item nicht verwendet werden. Auch mit Hilfe eines Items, welches z.B. auf einer 5-Punkte-Skala zu beantworten ist, können wohl kaum sehr viele Individuen in eine Rangreihe ohne Bindungen gebracht werden. Als Hauptproblem bei dieser Art der Skalenkonstruktion stellt sich damit das Finden geeigneter Items. Statt Items zu suchen, von denen jedes die N Individuen in eine Rangreihe bringt, kann die Rangreihe auch sukzessive erstellt werden. So wählt man zunächst ein Item (oder mehrere), mit dessen Hilfe die Population der N Individuen in wenige Gruppen geteilt wird, im Extremfall in nur zwei Gruppen. Im nächsten Schritt werden Items gewählt, die jede dieser Gruppen weiter unterteilen, usw., bis schließlich die N Individuen eine Rangreihe fast ohne Bindungen bilden. Diese Rangreihe liefert dann die Ausgangsdaten für das ordinale Testmodell. Hier findet auf jeder Stufe der Skalenkonstruktion eine Itemanalyse statt.

Das Kriterium für die Itemanalyse sind die Korrelationen zwischen der Rangreihe der Individuen, welche durch die Itembeantwortung entsteht, und der Rangreihe der Individuen bezüglich der Gesamtscores. Items mit niedrigen Korrelationen werden eliminiert.

Schließlich seien k'' Items ausgewählt worden, welche die Skala bilden.

3.2.2.3. Score des Individuums

Die so konstruierte Skala wird einer Eichpopulation von M Individuen vorgegeben. Diese Individuen werden bezüglich jedes Items in eine Rangreihe gebracht. Bei gleichen Werten können Mittelränge gebildet werden. Jedes Individuum erhält als Score die Summe seiner Rangzahlen zugeordnet. Diese Summenscores sind Ausgangsdaten für das ordinale Testmodell.

Die Summenscores werden in eine Rangreihe gebracht, wobei bei gleichen Scores jetzt die Rangzahlen zufällig zugeordnet werden. Hieraus wird pro Individuum der wahre Wert geschätzt, d.h.

$$\hat{T}_i = R_i, \quad i=1,\dots,M,$$

falls die M Individuen als Population aufgefaßt werden.

Damit diese Methode auch anwendbar ist, wenn nur ein einzelnes

Individuum vorhanden ist, kann mit Hilfe der Daten aus der Eichpopulation ein Manual erstellt werden. Hier wird angegeben, welche Scores pro Item zu welchen Rangzahlen führen. Zu jeder Rangsumme wird der geschätzte wahre Wert angegeben. Es kann zusätzlich der *relative Rang* angegeben werden, mit relativer Rang = Rangzahl/Populationsumfang.

3.2.2.4. Schätzen von Reliabilität und Validität

Soll die Reliabilität geschätzt werden, so wird eine weitere so konstruierte parallele Skala benötigt. Die Reliabilität kann dann durch die Korrelation zwischen den beiden parallelen Skalen geschätzt werden.
Eine "parallele Skala" ist nach Definition (Abschnitt I.1.5.) eine solche, die den Individuen dieselben wahren Werte zuweist wie die andere Skala. Man kann schon bei der Itemanalyse so viele Items auswählen, daß statt nur einer Skala zwei Skalen entstehen.
Die Validität kann geschätzt werden, wenn jedem Individuum zusätzlich noch eine Rangzahl bezüglich einer Kriteriumsvariablen zugewiesen wurde.
Die hier berechnete Reliabilität und Validität ist analog zu den entsprechenden Konzepten der KTT zu sehen.

3.2.2.5. Praktikabilität der beschriebenen Skalenkonstruktion

Die Praktikabilität dieser Art der Konstruktion einer Skala ist davon abhängig, ob geeignete Items gefunden werden können.

Ein gangbarer Weg scheint das angegebene sukzessive Vorgehen zu sein, bei dem die Probanden zunächst grob in wenige Gruppen geteilt werden und dann eine weitere Differenzierung innerhalb jeder Gruppe stattfindet.

3.3. Statistische Verfahren zum Vergleich von Rangreihen

Außer der Schätzung wahrer Werte von Objekten nach dem ordinalen Testmodell, der Schätzung der Reliabilität des beurteilenden Raters, usw., können auch Fragestellungen auftreten, die mit Hilfe statistischer Tests beantwortet werden sollen.

Beispielsweise wurde ein Markt- und Meinungsforschungsinstitut damit beauftragt, festzustellen, wie K verschiedene Werbespots in der Bevölkerung ankommen. Die Werbespots sollen von jedem urteilenden Individuum in eine Rangreihe gebracht werden. Es kann dann von Interesse sein, welche Rangreihe besonders häufig genannt wurde oder ob Werbespot i besonders häufig auf Rangplatz j genannt wurde oder ob im Mittel ein Unterschied in der Nennung der einzelnen Werbespots für die Rangplätze besteht.

Mit Hilfe dieser Ergebnisse sollen Entscheidungen getroffen werden. Ein Entscheidungskriterium ist z.B., denjenigen Werbespot auszuwählen, dessen Nennung für Rangplatz 1 signifikant ist.

Signifikanztests sind jedoch nur möglich mit der Nullhypothese der Gleichwahrscheinlichkeit der möglichen Rangreihen. Diese Hypothese bedeutet, daß sich die Werbespots in ihrer Wirkung nicht unterscheiden. Für das ordinale Testmodell ist dieses ein Spezialfall, der nicht betrachtet wird. Hier hätten alle Werbespots gleiche wahre Werte, was zur Folge hat, daß die Reliabilität Null ist.

Außer der Analyse einer Gruppe von Rangreihen kann auch die Fragestellung auftreten, zwei unabhängige Urteilergruppen miteinander zu vergleichen. Beispielsweise sollen die Werbespots in unterschiedlichen regionalen Gebieten beurteilt und die entstehenden Rangreihen verglichen werden.

Es werden im folgenden drei Möglichkeiten angegeben, mit Hilfe statistischer Tests Rangreihen miteinander zu vergleichen.
Lehmacher und Wall (1978) analysierten mit diesen Tests Stichproben von Verlaufskurven. In der hier betrachteten Situation werden statt der Verlaufskurven Rangreihen von Objekten betrachtet. Es wird angenommen, daß diese Objekte von mehreren Ratern je in eine Rangreihe gebracht wurden.

3.3.1. Vergleich zweier unabhängiger Gruppen von Ratern

Es werden zwei unabhängige Gruppen von Ratern bezüglich ihrer Beurteilung von K Objekten mit Hilfe statistischer Tests verglichen. Die hier betrachteten Tests arbeiten mit der Nullhypothese, daß die Verteilung der Rangreihen in beiden Gruppen identisch ist.

Jedes der Verfahren ist zweckmäßig für eine bestimmte inhaltliche Fragestellung. Um dieses anschaulicher zu gestalten, wird einmal angenommen, daß das Urteilsverhalten männlicher und weiblicher Rater bezüglich K Objekten miteinander verglichen werden soll. Dieser Vergleich kann sich beispielsweise auf die vollständigen Rangreihen beziehen. Will man wissen, ob in beiden Gruppen die Rangreihen unterschiedlich häufig genannt wurden, so ist ein (2×K!)-Felder-Test zu empfehlen. Hier werden die Rangreihen als Ganzes behandelt. Statt sich für die vollständigen Ranganordnungen zu interessieren, kann das Interesse auch dahingehend sein, daß man sich für Unterschiede in den Rangplätzen bestimmter Objekte interessiert. Beispielsweise kann die Fragestellung lauten: "Setzen die weiblichen Rater das Objekt Nr. 3 häufiger auf Rang 1 als die männlichen Rater?" In einem solchen Falle kann ein Anderson-Kannemann-Test durchgeführt werden. Das dritte Verfahren, welches hier vorgestellt wird, vergleicht die einzelnen Objekte bezüglich ihrer erhaltenen mittleren Rangzahlen. Es wird pro Objekt geprüft, ob ein Unterschied in der Beurteilung vorliegt. Diese Fragestellung ist gröber, sie bezieht sich nicht mehr auf einen einzelnen Rangplatz. Dieses kann zur Folge haben, daß möglicherweise Unterschiede im Urteilsverhalten der beiden Gruppen, die mit dem Anderson-Kannemann-Test entdeckt würden, hier nicht entdeckt werden.

3.3.1.1. (2×K!)-Felder-Test

Es kann ein Globaltest oder mehrere Einzeltests durchgeführt werden. Wie schon bemerkt wurde, kann dieser Test angewendet werden, um das Urteilsverhalten, hier also die Rangreihen der weiblichen und männlichen Rater, global miteinander zu vergleichen.

Das Vorgehen ist wie folgt:
Für die Rangreihen der beiden Gruppen wird eine Häufigkeitstabelle erstellt. Die Häufigkeiten seien mit m_k (für die 1. Gruppe mit M Ratern) bzw. mit n_k (für die 2. Gruppe mit N Ratern) bezeichnet. Die Rangreihen werden durchnumeriert, $k = 1,\ldots,K!$. Die Häufigkeitstabelle hat folgende Form (G1 = Gruppe 1, G2 = Gruppe 2):

Tab. 1.2.: Anzahlen der Rangreihen in den beiden Gruppen

Nr. der Rangreihe	G1	G2	
1	m_1	n_1	$l_1 = m_1 + n_1$
⋮	⋮	⋮	⋮
k	m_k	n_k	$l_k = m_k + n_k$
⋮	⋮	⋮	⋮
K!	$m_{K!}$	$n_{K!}$	$l_{K!} = m_{K!} + n_{K!}$
	M	N	M + N = L

Die Nullhypothese, daß die Rangreihen in beiden Gruppen identisch verteilt sind, läßt sich mit dem (2 ×K!)-Felder-Test (s. etwa Lienert, 1973, S. 182) überprüfen. Die Teststatistik lautet

$$\chi^2 = \sum_{k=1}^{K!} \left\{ \frac{(m_k - l_k \cdot \frac{M}{L})^2}{l_k \cdot \frac{M}{L}} + \frac{(n_k - l_k \cdot \frac{N}{L})^2}{l_k \cdot \frac{N}{L}} \right\} .$$

χ^2 ist asymptotisch χ^2-verteilt mit (K!-1) Freiheitsgraden.

Der Test hat den Nachteil, daß die Anzahl der Rangreihen mit K! wächst. Bei vielen zu beurteilenden Objekten und wenigen Ratern treten in der (2×K!)-Felder-Tafel vermehrt leere Zellen auf.

Man wird einen Einzeltest durchführen, wenn vor der Durchführung des Experimentes eine spezifizierte Hypothese vorliegt, z.B. daß sich die weiblichen und männlichen Rater bezüglich einer bestimmten Rangreihe Nr. k im Urteilsverhalten unterscheiden. Ein solcher Test wird als Vierfelder-Test durchgeführt. Die Vierfeldertafel lautet für die Rangreihe Nr. k:

Tab. 1.3.: Vierfeldertafel für den Vergleich der Häufigkeiten der Rangreihe Nr. k

	G1	G2	
Rangreihe Nr. k	m_k	n_k	l_k
alle anderen Rangreihen	$M-m_k$	$N-n_k$	$L-l_k$
	M	N	L

Die Auswertung kann asymptotisch mit einem χ^2-Test mit 1 Freiheitsgrad oder auch mit der Prüfgröße

$$X_k = \sqrt{L}\,\frac{[m_k(N-n_k) - n_k(M-m_k)]}{\sqrt{l_k(L-l_k)M\cdot N}}$$

geschehen, welche nach der Standardnormalverteilung beurteilt wird. Bei mehreren durchzuführenden Tests ist auf die Adjustierung des Signifikanzniveaus zu achten. Sollen alle Einzeltests durchgeführt werden, so handelt es sich um K! Vergleiche.

3.3.1.2. Vergleich zweier Anderson-Kannemann-Tafeln

Dieses Verfahren kann angewendet werden, wenn man sich speziell dafür interessiert, ob z.B. die weiblichen Rater das Objekt Nr. k mit einer größeren Wahrscheinlichkeit auf den Rangplatz Nr. 1 setzen als die männlichen Rater. Die Fragestellung bezieht sich auf ein (oder auch mehrere) Objekt. In 3.3.1.1 bezog sie sich auf eine Rangreihe. Somit wird die vorhandene Information etwas anders ausgenutzt. Es werden nicht, wie in 3.3.1.1, die vollständigen Rangreihen der beiden Gruppen verglichen. Für jede Gruppe wird die sogenannte Incidenztafel aufgestellt. Hierfür wird ausgezählt, wie häufig ein bestimmtes Objekt eine bestimmte Rangzahl erhalten hat. Die Incidenztafel der beiden Gruppen haben die Form:

Tab. 1.4.: Incidenztafeln der beiden Gruppen

Rang	Gruppe 1 Individuum 1 ... k ... K
1	$m_{11} \cdots m_{1k} \cdots m_{1K}$
⋮	⋮ ⋮ ⋮
l	$m_{l1} \cdots m_{lk} \cdots m_{lK}$
⋮	⋮ ⋮ ⋮
K	$m_{K1} \cdots m_{Kk} \cdots m_{KK}$

Rang	Gruppe 2 Individuum 1 ... k ... K
1	$n_{11} \cdots n_{1k} \cdots n_{1K}$
⋮	⋮ ⋮ ⋮
l	$n_{l1} \cdots n_{lk} \cdots n_{lK}$
⋮	⋮ ⋮ ⋮
K	$n_{K1} \cdots n_{Kk} \cdots n_{KK}$

Dabei gibt m_{11} (n_{11}) an, wie häufig Objekt 1 den Rangplatz 1 erhielt.

Unter der Nullhypothese haben die normierten Elemente der Incidenztafeln den gleichen Erwartungswert, d.h. es gilt

$$E[\frac{m_{lk}}{M}] = E[\frac{n_{lk}}{N}], \; l,k=1,\ldots,K \; .$$

Dieses kann für jedes Paar (l,k) mit einem Vierfelder-Test überprüft werden. Z.B. lautet die Vierfeldertafel für Objekt Nr. k und Rangplatz Nr. l:

Tab. 1.5.: Vierfeldertafel für Individuum Nr. k auf Rangplatz Nr. l

	G1	G2	
Rang l	m_{lk}	n_{lk}	$m_{lk}+n_{lk}$
andere Ränge	$M-m_{lk}$	$N-n_{lk}$	$L-m_{lk}-n_{lk}$
	M	N	L

Die Auswertung kann z.B. asymptotisch geschehen mit einem χ^2-Test oder mit der Prüfgröße (siehe Lehmacher und Wall, 1978; Lienert, 1978, S. 341):

$$X_{kl} = \sqrt{L}\,\frac{[m_{lk}(N-n_{lk})-n_{lk}(M-m_{lk})]}{\sqrt{m_{lk}+n_{lk})(L-m_{lk}-n_{lk})M\cdot N}} \; .$$

Diese ist nach der Standardnormalverteilung zu beurteilen.

Sollen mehrere Tests durchgeführt werden, muß das Signifikanzniveau adjustiert werden. Will man alle Einzelvergleiche durchführen, so sind dieses hier K^2, während es sich in 3.3.1.1. um K! Vergleiche handelte.

3.3.1.3. Vergleich zweier Friedman-Rangsummen-Vektoren

Für jede der Gruppen wird zunächst eine Friedman-Tafel aufgestellt. Die Friedman-Tafeln enthalten die von den Ratern genannten Rangreihen:

Tab. 1.6: Friedman-Tafeln für die beiden Gruppen

Rater	Individuum 1 ... k ... K
1	$R'_{11} \cdots R'_{1k} \cdots R'_{1K}$
⋮	⋮ ⋮ ⋮
m	$R'_{m1} \cdots R'_{mk} \cdots R'_{mK}$
⋮	⋮ ⋮ ⋮
M	$R'_{M1} \cdots R'_{Mk} \cdots R'_{MK}$
	$R'_{\cdot 1} \cdots R'_{\cdot k} \cdots R'_{\cdot K}$

Rater	Individuum 1 ... k ... K
1	$R_{11} \cdots R_{1k} \cdots R_{1K}$
⋮	⋮ ⋮ ⋮
n	$R_{n1} \cdots R_{nk} \cdots R_{nK}$
⋮	⋮ ⋮ ⋮
N	$R_{N1} \cdots R_{Nk} \cdots R_{NK}$
	$R_{\cdot 1} \cdots R_{\cdot k} \cdots R_{\cdot K}$

Die beiden Tafeln werden spaltenweise verglichen, d.h. es werden Vergleiche pro Objekt durchgeführt.

Für jedes k kann die Prüfgröße

$$V_k = \sqrt{\frac{(N+M-1)NM}{N+M}} \cdot \frac{(\bar{R}'_{\cdot k} - \bar{R}_{\cdot k})}{\sqrt{\sum_{i=1}^{M} R'^2_{ik} + \sum_{i=1}^{N} R^2_{ik} - \frac{1}{N+M}(R'_{\cdot k}+R_{\cdot k})^2}}$$

berechnet und asymptotisch nach einer Standardnormalverteilung beurteilt werden (z.B. Lienert 1978, S.339 oder Lehmacher und Wall, 1978).

Bei mehreren Tests ist auch hier das Signifikanzniveau zu adjustieren.
Insgesamt können K Vergleiche durchgeführt werden.

3.3.2. Analyse einer Gruppe von Ratern

Bei der Analyse einer Gruppe von Ratern handelt es sich nicht um eine Überprüfung der Homogenität der Rater. Da nicht bekannt ist, nach welcher Wahrscheinlichkeitsverteilung diese urteilen, ist die Homogenität nicht mit einem statistischen Test nachweisbar. Hier sollen die Rangreihen, welche von homogenen Ratern genannt wurden, näher untersucht werden. Beispielsweise werden bei der Beurteilung der Güte von Haushaltsgeräten, Urlaubsreisen, usw. homogene Beurteiler vorausgesetzt. Interessant ist die Analyse der genannten Rangreihen, die Plazierung der Objekte.
Soll z.B. untersucht werden, welche Rangreihen besonders häufig genannt werden, so kann ein Anpassungstest durchgeführt werden. Dieser arbeitet mit der Nullhypothese, daß die K! möglichen Rangreihen gleichwahrscheinlich sind, also alle etwa gleichhäufig genannt wurden. Ein signifikantes Ergebnis bedeutet eine Abweichung von dieser Gleichwahrscheinlichkeit. Welche Rangreihe dafür verantwortlich ist, kann nicht gesagt werden. Hierfür müssen Einzeltests durchgeführt werden.

Interessiert man sich nicht für die vollständigen Rangreihen, sondern dafür, ob eines der Objekte wesentlich häufiger als die anderen einen bestimmten Rangplatz erhält, so ist der Anderson-Kannemann-Test zu empfehlen. Beispielsweise kann die Frage auftreten, ob innerhalb der Gruppe der weiblichen Rater das zu beurteilende Objekt Nr. k wesentlich häufiger den 1. Rangplatz erhält. Bei dieser Art der Fragestellung wird man sich nicht für alle Rangplätze in gleicher Weise interessieren, sondern z.B. nur für den ersten, vielleicht noch für den letzten Rangplatz.

Der Friedman-Test, der hier in seiner ursprünglichen Form durchgeführt wird, prüft, ob die K Objekte im Mittel gleich beurteilt wurden.

3.3.2.1. Anpassungstest

Der Anpassungstest prüft, ob die K! möglichen Rangreihen mit der gleichen Wahrscheinlichkeit genannt wurden. Er arbeitet mit der Prüfgröße

$$X^2 = \sum_{k=1}^{K!} \frac{(n_k - \frac{N}{K!})^2}{N/K!} ,$$

welche asymptotisch χ^2-verteilt ist mit (K! - 1) Freiheitsgraden. n_k ist die Häufigkeit der Rangreihe Nr. k, N die Anzahl der Rater. Ein signifikantes Ergebnis sagt nichts darüber aus, welche Rangreihe (bzw. welche Rangreihen) wesentlich häufiger genannt wurde, als es unter der Nullhypothese zu erwarten wäre. Hierfür müßten wieder K! Einzeltests (für jede der K! Klassen) durchgeführt werden. Es handelt sich hier um Binomialtests mit $p_0 = 1/K!$ (siehe Lienert, 1973, S.143). Asymptotisch können die Einzeltests mit der Prüfgröße

$$X_k = (n_k - \frac{N}{K!}) \sqrt{\frac{N}{K!} (1 - \frac{1}{K!})} , \quad k=1,\ldots,K!$$

durchgeführt werden. X_k ist jeweils nach einer Standardnormalverteilung zu beurteilen.

Das Signifikanzniveau braucht nicht adjustiert zu werden, wenn eine gezielte Hypothese vorliegt, die z.B. in Voruntersuchungen gewonnen wurde oder aus der Literatur bekannt ist. Beispielsweise soll überprüft werden, ob die Rangreihe Nr. k überzufällig häufig genannt wurde. Dann ist nur ein Test durchzuführen.

3.3.2.2. Anderson-Kannemann-Test

Im Falle einer einzelnen Gruppe von Ratern wird nur eine Incidenztafel (vgl. Tab. 1.4.) nach Anderson-Kannemann aufgestellt (Anderson 1959, Schach 1976, Kannemann 1976). Unter der Nullhypothese beträgt der Erwartungswert für alle Elemente der Tafel jeweils $\frac{N}{K}$. Der Globaltest prüft, ob die K Objekte gleichwahrscheinlich auf den einzelnen Rangplätzen genannt wurden. Die Prüfgröße lautet

$$X^2_{AK} = \frac{K-1}{K} \sum_{l,k} \frac{(n_{lk} - \frac{N}{K})^2}{\frac{N}{K}} \quad .$$

n_{lk} gibt an, wie häufig Objekt k den Rang l erhielt.

X^2_{AK} ist asymptotisch χ^2-verteilt mit $(K-1)^2$ Freiheitsgraden.

Interessanter als der Globaltest, der über ein einzelnes Objekt nichts aussagt, sind Einzeltests. Es sind K^2 Einzeltests möglich, die für jedes der K^2 Elemente n_{lk} , l=1,...,K , k=1,...,K , der Inzidenztafel überprüfen, ob es dem Erwartungswert $\frac{N}{K}$ entspricht. Die Tests sind Binomialtests mit $p_o = \frac{1}{K}$. Asymptotisch erfolgt die Beurteilung wieder nach der Standardnormalverteilung. Die Prüfgrößen lauten

$$X_{lk} = (n_{lk} - \frac{N}{K}) \sqrt{\frac{N}{K} (1 - \frac{1}{K})} \quad , l=1,...,K \; , k=1,...,K \; .$$

Auch hier sollten spezifizierte Hypothesen vorliegen, um nicht K^2 Tests durchführen zu müssen. Werden mehrere Tests durchgeführt, so muß das Signifikanzniveau adjustiert werden.

3.3.2.3. Friedman-Test

Der Friedman-Test wird hier in seiner ursprünglichen Form durchgeführt (Friedman, 1937). Er untersucht anhand der Rangsummen, ob sich die Rater bezüglich ihrer Beurteilung der K Individuen unterscheiden. Der Test arbeitet mit der Prüfgröße

$$X^2_F = \frac{12}{NK(K+1)} \sum_{k=1}^{K} (R_{.k} - \frac{1}{2} N(K+1))^2 .$$

X^2_F ist asymptotisch χ^2-verteilt mit (K-1) Freiheitsgraden.

Auch hier kann man K Einzeltests durchführen mit den Prüfgrößen

$$X_k = \sqrt{12/[NK(K+1)]} \; (R_{.k} - N(K+1) \frac{1}{2}) , \; k=1,...,K \quad ,$$

welche asymptotisch nach einer Standardnormalverteilung zu beurteilen sind.

Da der Friedman-Test im Unterschied zum Anpassungstest und zum Anderson-Kannemann-Test mit den Rangsummen arbeitet, kann es sein, daß er Unterschiede in der Beurteilung der Objekte, die vom Anderson-Kannemann-Test und vom Anpassungstest entdeckt würden, nicht findet. Das ist z.B. der Fall, wenn bei 3 Objekten 50 Rater die Rangreihe (1,2,3) nennen und 50 Rater die Rangreihe (3,2,1). Die Friedman-Rangsummen sind für die 3 Objekte jeweils 200.

3.4. Beispiel

Bei dem Beispiel handelt es sich um das Beurteilen von Individuen. Häufig sind nicht nur die wahren Werte der Individuen und z.B. die Reliabilität der Beurteilung von Interesse, sondern es sind auch statistische Tests durchzuführen. Bei dem gewählten Demonstrationsbeispiel liegt eine solche kombinierte Fragestellung vor. Es wurde einer Untersuchung entnommen, in deren Rahmen 5 fiktive Bewerber für eine Hochschullehrerstelle in Allgemeiner Psychologie nach ihrer Eignung für diese Stelle in eine Rangreihe gebracht werden sollten. Die Beurteiler waren Psychologiestudenten. Sie wurden angewiesen, sich vorzustellen, sie seien Mitglieder der Berufungskommission. Da in Voruntersuchungen festgestellt wurde, daß sowohl das Geschlecht der Bewerber als auch das Geschlecht der Beurteiler einen Einfluß auf den Rangplatz haben, wurden für die Untersuchung zwei unabhängige Gruppen von Versuchspersonen (Urteilern) gebildet. Der einen Urteilergruppe wurden die Bewerber neutral, also ohne Angabe ihres Geschlechts, zur Beurteilung vorgelegt, der anderen Urteilergruppe mit Angabe des Geschlechts. Die globale Fragestellung der Untersuchung lautete: "Die Angabe des Geschlechts der Bewerber beeinflußt das Urteil". Die spezifische Fragestellung, welche auch mit einem statistischen Test überprüft wurde, bezog sich auf einen speziellen Bewerber und einen speziellen Rangplatz.

Der deskriptive Teil der Untersuchung bestand darin, jedem Bewerber einen wahren Wert zuzuweisen und die Güte der Urteile zu untersuchen. Die insgesamt 30 Urteiler (15 männlich, 15 weiblich) nannten für die Bewerber, welche mit B, C, N, P und R bezeichnet wurden, die folgenden Rangreihen:

Tab. 1.7.: Rangreihen pro Gruppe und nach Geschlecht der Rater

Gruppe 1 (ohne Angabe des Geschlechts der Bewerber)

männliche Rater						weibliche Rater					
Rater Nr.	Bewerber B	C	N	P	R	Rater Nr.	Bewerber B	C	N	P	R
1	5	2	1	4	3	8	4	2	3	5	1
2	3	2	1	5	4	9	4	3	1	5	2
3	4	3	1	5	2	10	4	2	1	5	3
4	2	3	1	5	4	11	2	1	4	5	3
5	3	2	1	5	4	12	4	3	1	5	2
6	2	4	3	5	1	13	5	1	3	4	2
7	4	2	1	5	3	14	4	1	3	5	2

Gruppe 2 (mit Angabe des Geschlechts der Bewerber)

männliche Rater						weibliche Rater					
Rater Nr.	Bewerber B	C	N	P	R	Rater Nr.	Bewerber B	C	N	P	R
15	4	1	2	5	3	23	4	2	1	5	3
16	5	4	2	1	3	24	4	1	2	5	3
17	4	3	2	5	1	25	4	2	1	5	3
18	5	1	2	4	3	26	5	1	2	4	3
19	4	3	1	5	2	27	4	3	1	5	2
20	4	1	3	5	2	28	3	2	1	5	4
21	4	3	1	5	2	29	5	1	3	4	2
22	3	1	2	5	4	30	4	1	3	5	2

Um die wahren Werte und die Varianzen für jeden Bewerber schätzen zu können, wird angenommen, daß die N=5 Bewerber die Population darstellen. Nach 2.1.1.1. wird bei g Ratern der wahre Wert geschätzt durch

$$\hat{T}_i = \frac{1}{g} \sum_{k=1}^{g} R_{ik}.$$

Die Varianz der Rangzahlen wird geschätzt durch

$$\hat{\sigma}_i^2 = \frac{1}{g-1} \sum_{k=1}^{g} (R_{ik} - \bar{R}_i)^2.$$

R_{ik} ist die Rangzahl des Bewerbers i in der Rangordnung des Raters k. Damit erhält man die folgende Tabelle:

Tab. 1.8: Geschätzte wahre Werte und Varianzen

		männliche Beurteiler					weibliche Beurteiler				
		B	C	N	P	R	B	C	N	P	R
G1	$\hat{T}_i$	3,3	2,6	1,3	4,9	3	3,9	1,9	2,3	4,9	2,1
	$\hat{\sigma}_i$	1,1	0,8	0,8	0,4	1,2	0,9	0,9	1,3	0,4	0,7
G2	$\hat{T}_i$	4,1	2,1	1,9	4,4	2,5	4,1	1,6	1,8	4,8	2,8
	$\hat{\sigma}_i$	0,6	1,2	0,6	1,4	0,9	0,6	0,7	0,9	0,5	0,7

Die männlichen Beurteiler nannten im Mittel den Bewerber N als besten Bewerber, die weiblichen Beurteiler den Bewerber C.

Als nächstes soll die Reliabilität der Rater geschätzt werden. Hierfür wird die Schätzfunktion

$$\hat{\rho}^2(R,T) = 1 - \frac{\frac{1}{N} \sum_i \mathrm{Var}_a(R_i)}{\mathrm{Var}_g(R)} = 1 - \frac{\frac{1}{N} \sum_i \mathrm{Var}_a(R_i)}{\frac{N^2-1}{12}}$$

benutzt (2.3.1.).

Tab. 1.9.: Geschätzte Reliabilität $\hat{\rho}^2(R,T)$

	G1	G2
m	0,59	0,51
w	0,60	0,76

In der Gruppe 1 unterscheidet sich die geschätzte Reliabilität der männlichen Beurteiler nicht von derjenigen der weiblichen Beurteiler. Anders verhält es sich in der Gruppe 2. Hier ist die geschätzte Reliabilität der weiblichen Beurteiler höher als die der männlichen Beurteiler. Ebenso kann man der Tabelle entnehmen, daß sich die geschätzte Reliabilität der männlichen Beurteiler in Gruppe 1 kaum vom Wert in Gruppe 2 unterscheidet, bei den weiblichen Beurteilern ist die geschätzte Reliabilität in Gruppe 2 höher als diejenige in Gruppe 1. Damit kann man zunächst sagen, daß die Angabe des Geschlechts der Bewerber sich bei den männlichen Beurteilern nicht auf die Urteilerübereinstimmung auswirkte. Bei den weiblichen Beurteilern fand durch die Angabe des Geschlechts eine Erhöhung der Urteilerübereinstimmung statt.
Betrachtet man z.B. die Rangreihen der männlichen Beurteiler in Gruppe 1, so stellt man fest, daß diese sich wohl einig sind in der Beurteilung der Bewerber N und P. Die schlechte Übereinstimmung wird z.B. von den Bewerbern B und R verursacht.

Nominierung des Bewerbers C auf Rangplatz 1

Die spezifische Fragestellung der Untersuchung bezog sich auf die Nominierung des Bewerbers C für den besten Rangplatz. Die geschätzten wahren Werte erscheinen für eine solche Betrachtung nicht geeignet. Sie stellen mittlere Rangzahlen dar. Von Interesse ist aber nur der Rang 1. Besser ist es, die Information in Form von Anderson-Kannemann-Tafeln zusammenzustellen (3.3.1.2.).

Tab. 1.10.: Anderson-Kannemann-Tafeln

Gruppe 1 (G1)

männliche Beurteiler

Rang	Bewerber B	C	N	P	R
1	0	0	6	0	1
2	2	4	0	0	1
3	2	2	1	0	2
4	2	1	0	1	3
5	1	0	0	6	0

weibliche Beurteiler

Rang	Bewerber B	C	N	P	R
1	0	3	3	0	1
2	1	2	0	0	4
3	0	2	3	0	2
4	5	0	1	1	0
5	1	0	0	6	0

Gruppe 2 (G2)

männliche Beurteiler

Rang	Bewerber B	C	N	P	R
1	0	4	2	1	1
2	0	0	5	0	3
3	1	3	1	0	3
4	5	1	0	1	1
5	2	0	0	6	0

weibliche Beurteiler

Rang	Bewerber B	C	N	P	R
1	0	4	4	0	0
2	0	3	2	0	3
3	1	1	2	0	4
4	5	0	0	2	1
5	2	0	0	6	0

Der Tabelle kann man entnehmen, daß in Gruppe 1 die männlichen Beurteiler den Bewerber C 0mal auf den 1. Platz setzten, die weiblichen Beurteiler hingegen 3mal für Bewerber C und Rangplatz 1 stimmten. In Gruppe 2 nannten die männlichen Beurteiler den Bewerber C 4mal für den 1. Rangplatz, die weiblichen Beurteiler ebenfalls. Damit unterscheiden sich die weiblichen Beurteiler in Gruppe 2 kaum von den weiblichen Beurteilern in Gruppe 1. Bei den männlichen Beurteilern liegt jedoch ein Unterschied vor. Diese spezielle Information konnte z.B. aus den geschätzten wahren Werten nicht entnommen werden. Führt man für Bewerber C und Rangplatz 1 einen Einzeltest durch, so erhält man

Tab. 1.11.: Vierfeldertafel für Bewerber C

	G1	G2	
Rang 1	0	4	4
andere Ränge	7	4	11
	7	8	15

Unter der Nullhypothese, daß die beiden Gruppen gleich sind, ist die Überschreitungswahrscheinlichkeit im exakten Fisher-Test $p = 0.051$. Damit kann gesagt werden, daß bei diesem speziellen Problem das Merkmal "Geschlecht der Bewerber" einen Einfluß auf das Urteil der männlichen Rater hatte. Die weiblichen Beurteiler waren weniger zu beeinflussen.

Damit ist das Beispiel nicht nur ein Beispiel für die Anwendung des ordinalen Testmodells beim Skalieren von Individuen, sondern zeigt auch, daß männliche und weibliche Rater unterschiedlich urteilen können, also insgesamt keine homogene Ratergruppe bilden. Für die Anwendung scheint eine ausführliche Rateranalyse unerläßlich zu sein.

II. Verwendung Log-linearer Modelle zur Skalierung

In diesem Abschnitt soll überlegt werden, ob Log-lineare Modelle, die aus der Analyse von Kontingenztafeln bekannt sind, als Skalierungsmodelle verwendet werden können. Damit sollen Modelle, die im Bereich der Statistik Verwendung finden, bezüglich ihrer Verwendbarkeit in der Psychometrie diskutiert werden. Um Unterschiede und Gemeinsamkeiten dieser beiden Einsatzgebiete feststellen zu können, erfolgt eine kurze Einführung in die Theorie der Log-linearen Modelle. Diese Einführung kann angesichts des umfangreichen Gebietes nur kurz sein. Beispielsweise wird auf eine Darstellung der unterschiedlichen Arten der Datenerhebung, sowie der Parameterschätzung und Modellanpassung verzichtet, obwohl dieses für eine konkrete Anwendung wichtig ist. Ausführliche Darstellungen der Theorie Log-linearer Modelle finden sich z.B. in
Bishop, Fienberg und Holland (1975), Fienberg (1977), Haberman (1974), Darroch und Ratcliff (1972), Upton (1978), Fahrmeir und Hamerle (1984).

1. Log-lineare Modelle

Um besser verstehen zu können, was die Modelle im Rahmen der Analyse von Kontingenztafeln leisten, sei ihre historische Entwicklung betrachtet. Die ersten wichtigen Ergebnisse in der Analyse kategorieller Daten stammen von Pearson (1900) und Yule (1900) um die Jahrhundertwende.
So wurden beispielsweise der χ^2-Test auf Unabhängigkeit in 2-dimensionalen Kontingenztafeln entwickelt und verschiedene Koeffizienten zum Messen der Assoziation in Kontingenztafeln definiert. Wichtig für die Entwicklung der Log-linearen Modelle ist von diesen Koeffizienten das Kreuzproduktverhältnis, welches auch als "odds ratio" bezeichnet wird. Es wurde von Yule (1900) als Maß der Assoziation in 2×2-Tafeln definiert und lautet

$$C = \frac{p_{11}p_{22}}{p_{21}p_{12}} ,$$

wobei die p_{ij}, i,j=1,2 die Zellwahrscheinlichkeiten in einer Vierfeldertafel sind.

Die Ergebnisse wurden zunächst nicht auf mehrdimensionale Tafeln verallgemeinert. Mehrdimensionale Kontingenztafeln wurden in der Regel analysiert, indem man die Randtafeln analysierte. Interaktionen höherer Ordnung konnten so nicht getestet werden.
Erst Bartlett (1935) definierte unter Verallgemeinerung des odds ratio einen Koeffizienten zum Erfassen der Interaktion 2. Ordnung in 2 × 2 × 2-Tafeln.
Roy und Kastenbaum (1956) wiederum verallgemeinerten diesen Koeffizienten auf R × S × T-Tafeln, um eine Interaktion 2. Ordnung testen zu können. Die Hypothese, daß keine Interaktion 2. Ordnung vorliegt, lautet hiernach

$$\frac{p_{RST}\, p_{ijT}}{p_{iST}\, p_{RjT}} = \frac{p_{Rjk}\, p_{ijk}}{p_{iSk}\, p_{Rjk}} \quad \text{, mit } 1 \leq i \leq R-1,\ 1 \leq j \leq S-1,\ 1 \leq k \leq T-1 \text{ .}$$

p_{ijk} ist die Zellwahrscheinlichkeit für die Zelle (i,j,k).

Birch (1963) schrieb diese Bedingung um. Mit

$$p_{ijk} = \frac{m_{ijk}}{m_{...}} \text{ ,}$$

wobei m_{ijk} der Erwartungswert der Zelle (i,j,k) und $m_{...}$ die Gesamtzahl der Beobachtungen in der Kontingenztafel ist, erhält man durch Logarithmieren und Auflösen nach $\log m_{ijk}$ eine Gleichung, die linear in den Logarithmen ist:

$$\log m_{ijk} = \log m_{RST} - \log m_{iST} - \log m_{RjT} - \log m_{RSk} +$$
$$+ \log m_{ijT} - \log m_{ijk} - \log m_{Rjk}$$

mit $1 \leq i \leq R-1,\ 1 \leq j \leq S-1,\ 1 \leq k \leq T-1$.

Die Modellgleichung kann wie folgt geschrieben werden:

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} \text{ ,}$$

mit $1 \leq i \leq R,\ 1 \leq j \leq S,\ 1 \leq k \leq T$.

Diese Darstellung wird als *Log-lineares Modell* bezeichnet. Das allgemeine Log-lineare Modell enthält noch einen Interaktionsterm 2. Ordnung $u_{123(ijk)}$. Mit Hilfe Log-linearer Modelle wurde eine vollständige Assoziationsanalyse mehrdimensionaler Kontingenztafeln möglich.
Das Modell enthält für jede Zelle der Tafel einen Assoziationsparameter. Diese sind jedoch nicht unabhängig voneinander. Die Assoziation zwischen den Variablen der Kontingenztafel wird auf spezielle Art und Weise definiert, und zwar auf der Basis der odds ratios der Zellwahrscheinlichkeiten bzw. der Erwartungswerte der Zellhäufigkeiten. Ein Log-lineares Modell kann für die Erwartungswerte der Zellhäufigkeiten oder für die Zellwahrscheinlichkeiten formuliert werden. Enthält das Modell genauso viele unabhängige Parameter wie die Kontingenztafel Zellen, so wird es *saturiertes* Modell genannt. Ist die Zahl der Parameter geringer als die Zahl der Zellen, so spricht man von einem *nichtsaturierten* Modell. Der Einfachheit halber soll nicht das oben angegebene Modell für eine dreidimensionale Kontingenztafel betrachtet werden, sondern ein Modell für eine zweidimensionale Tafel mit R Ausprägungen der Zeilenvariablen (i=1,...,R) und J Ausprägungen der Spaltenvariablen (j=1,...,J). Es kann das folgende saturierte Log-lineare Modell angenommen werden (Bishop, Fienberg, Holland, 1975, S. 24):

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)} \quad , \; i=1,...,R, \; j=1,...,J$$

mit den Restriktionen

$$\sum_i u_{1(i)} = \sum_j u_{2(j)} = \sum_i u_{12(ij)} = \sum_j u_{12(ij)} = 0 \quad .$$

Das Modell besagt, daß sich der Logarithmus des Erwartungswertes der Zellhäufigkeiten additiv aus einem von i und j unabhängigen Parameter, aus Parametern, die jeweils nur von einer Variablen der Kontingenztafel abhängen, sowie aus Interaktionsparametern zusammensetzt. Die Parameter sind leicht zu interpretieren. u ist der Gesamtmittelwert der Logarithmen:

$$\sum_{i,j} \log m_{ij} = \sum_{i,j} u = R \, J \, u \quad ,$$

$$u = \frac{1}{R\,J} \sum_{i,j} \log m_{ij} \ .$$

Die $u_{1(i)}$, i=1,...,R, werden als "Zeileneffekte" bezeichnet, mit

$$u_{1(i)} = \frac{1}{J} \sum_{j} \log m_{ij} - u.$$

Die $u_{2(j)}$, j=1,...,J, werden als "Spalteneffekte" bezeichnet, mit

$$u_{2(j)} = \frac{1}{R} \sum_{i} \log m_{ij} - u.$$

$u_{12(ij)}$ sind die Interaktionsparameter für jede der RJ Zellen. Es ist

$$u_{12(ij)} = \log m_{ij} - \frac{1}{J} \sum_{j} \log m_{ij} - \frac{1}{R} \sum_{i} \log m_{ij} + u.$$

Ist $u_{12(ij)}=0$, so sind die beiden Merkmale voneinander unabhängig. Man erhält dann das sogenannte Unabhängigkeitsmodell

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} \ .$$

Durch die Bezeichnung "Zeileneffekte" und "Spalteneffekte" wird eine Analogie zur Varianzanalyse nahegelegt. Fienberg (1977, S. 4) bemerkt, daß dieses nur gerechtfertigt ist, wenn z.B. bei einer zweidimensionalen Kontingenztafel eine Variable eine Einflußvariable und die andere Variable eine Antwortvariable ist und wenn man sich nur für den Effekt der Einflußvariablen (unabhängige Variable) auf die Antwortvariable (abhängige Variable) interessiert. Liegt eine solche asymmetrische Fragestellung vor, so kann sie direkt im Modell berücksichtigt werden . Dieses leistet ein Logit-Modell (Bishop, Fienberg und Holland, 1975, S. 30).

Wenn es sich um eine R × J-Kontingenztafel handelt, können *Logits* auf unterschiedliche Art und Weise gebildet werden, z.B.

"akkumulierte Logits": $L_{ij} = \log \left(\frac{m_{i,j+1} + \ldots + m_{iJ}}{m_{i1} + \ldots + m_{ij}} \right)$,

"continuation ratio"-Logits: $L_{ij} = \log \left(\frac{m_{i,j+1}}{m_{i1} + \ldots + m_{ij}} \right)$

"aufeinanderfolgende Kategorien"-Logits: $L_{ij} = \log \left(\frac{m_{i,j+1}}{m_{ij}} \right)$.

Ein mögliches Logit-Modell lautet

$$\log \left(\frac{m_{i,j+1}}{m_{ij}} \right) = w_{1(i)} + w_{2(j)}$$

mit $\sum_i w_{1(i)} = 0$.

Log-lineare und Logit-Modelle sind ineinander überführbar (siehe hierzu auch Goodman, 1983).

Beispielsweise kann das ebengenannte Modell durch Subtraktion aus einem Log-linearen Modell entstanden sein:

$$\log m_{i,j+1} - \log m_{i,j} = (u_{2(j+1)} - u_{2(j)}) + (u_{12(i,j+1)} - u_{12(i,j)}) =$$

$$= w_{2(j)} + w_{1(i)} .$$

Die anderen Parameter verschwinden durch die Subtraktion. Wie man sieht, ist $w_{1(i)}$ der Parameter, welcher den Effekt der unabhängigen Zeilenvariablen auf das Verhältnis $\frac{m_{i,j+1}}{m_{i,j}}$ angibt.

In den Modellen, welche bisher erwähnt wurden, werden die Merkmale so behandelt, als hätten sie Nominalskalenniveau. D.h. die Kategorien können vertauscht werden, ohne daß sich z.B. Parameterschätzungen oder χ^2-Statistiken ändern. Ausgehend von diesen Modellen wurden in der letzten Zeit spezielle Modelle entwickelt, welche ordinale Merkmale berück-

sichtigen. McCullagh (1980) betont, daß hier die in der Kontingenztafel betrachteten Merkmale deutlich nach Antwortvariablen und Einflußfaktoren unterschieden werden sollten.

Die Modelle für Kontingenztafeln mit geordneten Kategorien unterteilt Agresti (1983) in drei Gruppen:

1. Log-lineare Modelle,
2. Logit-Modelle,
3. Alternative Modelle.

In einer zweidimensionalen Kontingenztafel können entweder eine oder zwei der Variablen ordinal sein. Man erhält so eine nominal-ordinal-Tafel bzw. eine ordinal-ordinal-Tafel. Z.B. soll angenommen werden, daß nur die Spaltenvariable, welche die Antwortvariable darstellen soll, ordinal ist. Die Zeilenvariable stellt den Einflußfaktor dar. In Log-linearen Modellen, welche den ordinalen Charakter der Antwortvariablen berücksichtigen sollen, werden üblicherweise den geordneten Kategorien Zahlen zugeordnet: $v_1 < v_2 < \ldots < v_J$ (Fienberg, 1977, S. 62). Diese Zahlen sollen keine Skalenwerte für die Kategorien darstellen. Sie haben nur die Funktion, den ordinalen Charakter der Spaltenvariablen ins Modell zu bringen (siehe Goodman 1983, Agresti, 1983). Ein Log-lineares Modell für eine nominal-ordinal-Tafel ist z.B. (Agresti, 1983):

$$\log m_{ij} = u+u_{1(i)}+u_{2(j)}+\tau_{1(i)}(v_j-\bar{v}),\ i=1,\ldots,R,\ j=1,\ldots,J,$$

$$\text{mit } \sum_i u_{1(i)} = \sum_j u_{2(j)} = \sum_i \tau_{1(i)} = 0 .$$

Das Modell fordert eine bestimmte Interaktionsstruktur, nämlich $u_{12(ij)} = \tau_{1(i)}(v_j-\bar{v})$. Will man die Zuordnung von Zahlen $v_1 < v_2 < \ldots < v_J$ zu den geordneten Kategorien vermeiden, so ist ein Logit-Modell vorzuziehen, z.B.

$$L_{ij} = \log\left(\frac{m_{i,j+1}+\ldots+m_{iJ}}{m_{i1}+\ldots+m_{ij}}\right) = u_j+u_{1(i)},\ i=1,\ldots,R,\ j=1,\ldots,J-1,$$

$$\text{mit } \sum_i u_{1(i)} = 0 .$$

Zu jeder Zeile i gehören (J-1) Logits. Diese werden mit Hilfe der kumulierten Erwartungswerte (bzw. Wahrscheinlichkeiten) gebildet. Durch die Verwendung kumulierter Wahrscheinlichkeiten wird der ordinale Charakter der Spaltenvariablen berücksichtigt.
Log-lineare und Logit-Modelle der unterschiedlichsten Art betrachten z.B. Goodman (1979), Clayton (1974), Andrich (1979), Haberman (1974) und Agresti und Kezouh (1983). Goodman (1983) betrachtet ausführlich verschiedene Log-lineare Modelle und ihre Äquivalenz zu Logit-Modellen.

Zu den "alternativen Modellen" zählt Agresti (1983) Regressionsmodelle. Für eine ordinal-ordinal-Tafel schlägt er das folgende Modell vor:

$$M_i = \mu + \beta^x (u_i - \bar{u}), \quad i=1,\ldots,R.$$

Dabei werden den Kategorien der beiden ordinalen Variablen die Zahlen $u_1 < u_2 < \ldots < u_R$ und $v_1 < \ldots < v_J$ zugewiesen und es ist

$$M_i = \sum_j v_j m_{ij}/n_{i.} \quad , \quad i=1,\ldots,R.$$

β^x ist der Regressionskoeffizient.

Anderson und Philips (1981) schlagen ebenfalls Regressionsmodelle vor. Es wird eine stetige latente Variable angenommen, die nur ordinal beobachtbar ist. Werte für die latente Variable sollen aus bekannten Prädiktorvariablen vorhergesagt werden. In Anlehnung an Plackett (1974, S. 66) und an McCullagh (1980) wird das folgende Modell angegeben:

$$P(Y \leq s | \underset{\sim}{x}) = \frac{\exp(\underset{\sim}{\theta}_s - \underset{\sim}{\beta}^T \underset{\sim}{x})}{1+\exp(\underset{\sim}{\theta}_s - \underset{\sim}{\beta}^T \underset{\sim}{x})} \quad , \quad s=0,1,\ldots,K \ .$$

Y stellt die ordinale Antwortvariable dar und $\underset{\sim}{x} = (X_1,\ldots,X_p)$ die p Prädiktorvariablen.
Für die Komponenten des unbekannten Parametervektors $\underset{\sim}{\theta}_s$ gilt $\theta_1 < \theta_2 < \ldots < \theta_{K-1}$, $\theta_0 = -\infty$, $\theta_K = +\infty$. Der Vektor $\underset{\sim}{\beta}^T = (\beta_1,\ldots,\beta_p)$

repräsentiert den unbekannten "Regressionskoeffizienten".

Die Regressionsmodelle sind nur ein Beispiel der sogenannten alternativen Modelle (Agresti, 1983). Erwähnt werden sollen noch Semenya und Koch (1979), welche den Kategorien des Einflußfaktors mittlere Ridit-Scores zuordnen und Williams und Grizzle (1972), welche Perzentilscores zuweisen. Semenya et al. (1983) bilden zum Vergleich zweier Kategorien i und i' des Einflußfaktors die Statistik

$$U_{ii'} = \sum_{j=1}^{r} (g_{i'j} - \frac{1}{2} p_{i'j}) p_{ij} \cdot$$

Dabei ist $p_{ij} = \frac{n_{ij}}{n_{i\cdot}}$ die Schätzung der Zellwahrscheinlichkeit in der i-ten Kategorie, $g_{i'j} = \sum_{j'=1}^{j} p_{i'j'}$, die kumulierte Häufigkeit. Die Schätzung wird zum Vergleich je zweier Kategorien verwendet. Sie wird von den Autoren z.B. auf das Bradley-Terry Modell angewendet.

Zusammenfassend kann zum Einsatz von Log-linearen und Logit-Modellen in der Analyse mehrdimensionaler Kontingenztafeln gesagt werden, daß mit ihrer Hilfe die Struktur zwischen den Variablen, d.h. die Interaktionen, beschrieben werden. Die Variablen selbst finden im Rahmen der Analyse von Kontingenztafeln keine große Beachtung. Während Pearson (1900) annahm, daß hinter den Kategorien der Kontingenztafel grundsätzlich stetige Variablen stehen, nahm Yule (1900) die Gegenposition zu Pearson ein. Er nahm keine latenten Variablen an, die "hinter den Kategorien" stehen, sondern untersuchte strukturelle Beziehungen zwischen den in der Kontingenztafel gegebenen Variablen. Fienberg (1977, S. 5) bezeichnet ihn deshalb als den Begründer der Log-linearen Modellschule in der Kontingenztafelanalyse. In der statistischen Literatur der letzten 30 Jahre herrscht die Auffassung Yules vor. Somit hat vor der Erstellung einer Kontingenztafel i.a. keine Messung oder Skalierung stattgefunden. Setzt die Anwendung eines Modells Zahlen voraus, so

werden diese den Kategorien häufig mehr oder weniger willkürlich zugewiesen. Dieses hat zur Folge, daß z.B. Regressionsmodelle, welche intervallskalierte Prädiktoren voraussetzen, auch angewendet werden, wenn die Variablen nur ordinal- oder nominalskaliert sind. Damit wird deutlich, daß Messen bzw. Skalieren, welches der Anwendung von Statistik vorgeordnet ist, auch im Rahmen der Analyse von Kontingenztafeln eine wichtige Rolle spielt.

2. Log-lineare Modelle und Skalierung

Die Zielsetzungen von Kontingenztafelanalyse und Skalierung sind unterschiedlich. Bei der Analyse von Kontingenztafeln ist die Hauptfragestellung i.a. die Untersuchung des Zusammenhanges von Variablen. Mit Skalierungs- bzw. Meßmodellen sollen psychologische Merkmale auf einem geeigneten Skalenniveau erfaßt werden. Sowohl die Untersuchung des Zusammenhanges von Variablen wie auch deren Skalierung können gemeinsam im Rahmen einer Fragestellung auftreten. In der Literatur zur Analyse von Kontingenztafeln findet man Beispiele, in denen zweifellos beide Aspekte vorkommen, der Skalierungsaspekt jedoch nicht erwähnt wird. Zusammenhangsanalyse wie auch Skalierung könnten mit demselben Modell durchgeführt werden. Im folgenden seien drei dieser Beispiele aus der Literatur erwähnt, nämlich Simon (1974), Williams und Grizzle (1972) und Yates (1948).

In Simons Beispiel beantworten Individuen aus R Personengruppen eine bestimmte Frage auf einer 5-Punkte-Skala. Die R Personengruppen stellen die Ausprägungen der Einflußvariablen dar, die 5-Punkte-Skala die Zielvariable mit geordneten Kategorien. Für die Assoziationsanalyse der Kontingenztafel gibt Simon (1974) zwei Logit-Modelle an.

Bei Williams und Grizzle (1972) findet man eine ähnliche Situation. Die Autoren betrachten eine dreidimensionale Kontingenztafel mit den beiden Einflußfaktoren "Urbanisierung" und "Ort". Die ordinale Antwortvariable ist eine 9-Punkte-Kategorienskala. Auf dieser soll die Präferenz für schwarze Oliven angegeben werden. Die Autoren schlagen zur Assozia-

tionsanalyse ein Logit-Modell vor.
Yates (1948) betrachtet Kinder, welche nach 5 unterschiedlichen Hausarbeitsmethoden und einem Lehrerrating klassifiziert werden. Die Lehrer beurteilen auf einer geordneten 3-Punkte-Skala den Erfolg der Hausarbeit.

In allen Beispielen werden die Effekte der Einflußfaktoren auf die ordinalen Antwortvariablen geschätzt und auf Signifikanz getestet.
Für die praktische Anwendung ist möglicherweise die Signifikanz von Effekten zweitrangig. Interessanter wird die Frage sein, welche der Personengruppen die gestellte Frage in einer niedrigen Kategorie beantworten, welche in einer hohen Kategorie (Simon, 1974). In Beispiel 2 wird der Auftraggeber der Untersuchung über schwarze Oliven erfahren wollen, in welchem Ort eine große Vorliebe für schwarze Oliven besteht und in welchem Ort die Vorliebe geringer ist. Dieses ist keine Frage der Signifikanz von Effekten, sondern der Skalierung der Orte nach ihrer Präferenz für schwarze Oliven. Auch beim letzten Beispiel wird es interessant sein zu wissen, welche Methode der Hausarbeit am effektivsten war.

Bei den drei Beispielen handelte es sich jeweils um eine asymmetrische Fragestellung, d.h. die Variablen werden nach Einflußfaktoren und Zielvariable unterschieden. Die Ausprägungen der nominalen Einflußvariablen waren die Objekte (Personengruppen, Orte, Hausarbeitsmethoden). Die ordinale Antwortvariable war eine geordnete Kategorienskala, auf welcher die Objekte beurteilt wurden. Die Beispiele lassen vermuten, daß Log-lineare- bzw. Logit-Modelle mit einer ordinalen Antwortvariablen zur Skalierung verwendet werden können. Im folgenden Abschnitt wird ein spezielles Logit-Modell bezüglich seiner Eigenschaften als Meßmodell untersucht. Latente Variablen werden nicht angenommen.

Ein Hinweis bezüglich der Verwendung von Regressionsmodellen zur Skalierung findet sich bei Anderson und Philips (1981). Die Autoren empfehlen, eine als stetig vorausgesetzte latente Variable mit einem ordinalen Regressionsmodell zu skalieren. Da im Rahmen der vorliegenden Arbeit keine stetigen

latenten Variablen vorausgesetzt werden, wird diese Möglichkeit nicht näher betrachtet.
Auch Hinweise über die Äquivalenz gewisser Log-linearer Modelle und probabilistischer Meßmodelle finden sich in der Literatur.

Mellenbergh und Vijn (1981) stellen das Rasch Modell als Log-lineares Modell dar, Perline, Wright und Wainer (1979) präsentieren das Rasch Modell als additive conjoint measurement, Brogden (1977) verbindet Rasch Modell, Law of comparative judgement und additive conjoint measurement.

Kelderman (1984) formuliert das Rasch Modell als Quasi-Unabhängigkeitsmodell. Log-lineare Modelle bzw. deren multiplikative äquivalente Formulierungen wurden ebenfalls verwendet, um Anwortmuster zu skalieren, was bei der Guttman-Skala zur Anwendung kommt (Davison, 1980, Goodman, 1975, Fienberg 1977, S. 155).

Fienberg und Larntz (1976) geben eine Darstellung des Bradley-Terry Modells als Log-lineares Modell (siehe hierzu auch Fienberg 1977, S. 150, Fienberg, 1979). Auch Semenya u.a. (1983) geben eine Darstellung des Bradley-Terry Modells als Logit-Modell. Damit wird eine Analogie hergestellt zwischen probabilistischen psychometrischen Skalierungsmodellen und Log-linearen Modellen.

Log-lineare Modelle arbeiten i.a. ohne die Annahme latenter Variablen. Weiterhin ist es erforderlich, daß die Daten sich in Form einer Kontingenztafel darstellen lassen. Psychometrische Skalierungsmodelle arbeiten i.a. mit der Annahme latenter Variablen, und die Daten sind nicht immer in Form einer Kontingenztafel darstellbar. Die erwähnte Analogie ist möglicherweise nur formaler Natur und soll am Beispiel des Rasch Modells und des BTL-Modells betrachtet werden.

2.1. Skalierung von Objekten (Items) mit einem speziellen Logit-Modell

Es werde angenommen, daß R Objekte (Items) gegeben sind, die von g homogenen Ratern bezüglich ihrer Merkmalsausprägung auf einer geordneten K-Punkte-Skala beurteilt werden. Es wird weiterhin angenommen, daß die Rater unabhängig voneinander urteilen und daß jeder Rater die Objekte unabhängig voneinander beurteilt. Es entsteht pro Objekt eine Häufigkeitsverteilung der Urteile über den Kategorien. Diese Häufigkeitsverteilungen kann man in einer Kontingenztafel mit festen Zeilensummen darstellen:

Tab. 2.1.: Kontingenztafel der Urteilsdaten

Objekte	1	...	j	...	K	
O_1	n_{11}	...	n_{1j}	...	n_{1K}	$n_{1.}=g$
O_i	n_{i1}	...	n_{ij}	...	n_{iK}	$n_{i.}=g$
O_R	n_{R1}	...	n_{Rj}	...	n_{RK}	$n_{R.}=g$
	$n_{.1}$	...	$n_{.j}$	...	$n_{.K}$	R·g

In jeder Zeile der Kontingenztafel stehen g Urteile, insgesamt enthält die Tafel R·g Urteile. Die Randsummen $n_{i.}$, $i=1,...,R$, sind fest. n_{ij} ist die Anzahl der Rater, welche das Objekt i in die Kategorie j eingeordnet haben. Wäre die Urteilsverteilung der Rater eine Einpunktverteilung, so bestünde die Kontingenztafel fast nur aus leeren Zellen. Ein solcher Fall wird im Modell ausgeschlossen.

Bei der so entstehenden Kontingenztafel werden die Ausprägungen der Zeilenvariablen durch die Objekte repräsentiert. Diese Variable wird zunächst als nominal skaliert aufgefaßt. Die Spaltenvariable ist eine ordinale Antwortvariable mit K Ausprägungen. Die vorliegende Kontingenz-

tafel ist vom Typ nominal-ordinal.

Im folgenden wird ein spezielles Logit-Modell daraufhin untersucht, ob es zur Skalierung der Objekte verwendet werden kann. Ein Logit-Modell erscheint angemessen, da es sich um eine nominal-ordinal-Tafel handelt und weiterhin eine Randsumme fest ist (Bishop u.a., 1975, S. 30). Die Logits berücksichtigen den ordinalen Charakter der Antwortvariablen. Die Zuordnung von Zahlen zu den geordneten Kategorien wird vermieden.

Modell

Für die Zellen (ij), i=1,...,R, j=1,...,K-1, der Kontingenztafel werden "akkumulierte" Logits gebildet:

$$L_{ij} = \log \left(\frac{m_{i,j+1}+\ldots+m_{iK}}{m_{i1}+\ldots+m_{ij}} \right) ,$$

m_{ij} ist der Erwartungswert für die Zelle (i,j).

Mit den L_{ij} wird ein einfaches additives Logit-Modell formuliert (Agresti, 1983):

$$L_{ij} = \mu_j + \tau_i^x \quad , \; i=1,\ldots,R, \; j=1,\ldots,K-1, \quad \sum_i \tau_i^x = 0 \; .$$

Das Modell ist gleichbedeutend damit, daß die Verteilungsfunktion $F_i(j)$ der Zufallsvariablen

Y_i = Beurteilung des Objektes i auf der ordinalen K-Punkte-Skala

eine bestimmte Form hat:

$$L_{ij} = \log \frac{1-F_i(j)}{F_i(j)} = \mu_j + \tau_i^x \; ,$$

$$F_i(j) = \frac{1}{1+e^{\mu_j+\tau_i^x}} \; .$$

Das oben genannte Logit-Modell soll genauer betrachtet werden: Für jede Zeile i erhält man (K-1) Logits. Diese sind um so

größer, je größer die Erwartungswerte in den höheren Kategorien sind, je weiter rechts also die Wahrscheinlichkeitsmasse liegt.
Die Parameter μ_j und τ_i^x sind wie folgt zu interpretieren:
Die Summation der Modellgleichung über i ergibt

$$\sum_i L_{ij} = R\,\mu_j \quad ,$$

damit ist

$$\mu_j = \frac{1}{R} \sum_i L_{ij} \quad .$$

μ_j ist der Mittelwert der Logits über die Zeilen für die feste Spalte j. Summiert man statt über i jetzt über j, so ergibt sich:

$$\sum_j L_{ij} = \sum_j \mu_j + \sum_j \tau_i^x = \frac{1}{R} \sum_j \sum_i L_{ij} + (K-1)\,\tau_i^x \quad ,$$

$$\tau_i^x = \frac{1}{(K-1)} \sum_j L_{ij} - \frac{1}{R(K-1)} \sum_{i,j} L_{ij} \quad .$$

Bezeichnet man mit

$$\frac{1}{R(K-1)} \sum_{i,j} L_{ij} = \mu_{..}$$

den Mittelwert der insgesamt R(K-1) Logits,
so ergibt sich für τ_i^x:

$$\tau_i^x = \frac{1}{K-1} \sum_j L_{ij} - \mu_{..} \quad .$$

Sieht man einmal von $\mu_{..}$ ab, dann wird τ_i^x von der Summe der Logits in Zeile i bestimmt. Die Summe ist um so größer, je größer die Erwartungswerte in den höheren Kategorien sind, je weiter rechts also die Wahrscheinlichkeitsmasse der Verteilung liegt. Damit ist auch τ_i^x um so größer, je weiter rechts die Wahrscheinlichkeitsmasse der Verteilung liegt. τ_i^x wird als "Zeileneffekt" bezeichnet. Der Parame-

Objekt) auf die ordinale Antwortvariable (das Urteil des Raters) ist.

Als nächstes werden zwei beliebige Objekte O_i und $O_{i'}$, also zwei Zeilen i und i' , betrachtet.
Bildet man die Differenz der Logits, so ergibt sich

$$L_{ij}-L_{i'j} = \tau_i^x - \tau_{i'}^x \ .$$

Die Differenz ist unabhängig von j. Damit muß für alle j=1,...,K-1 gelten

$$L_{ij}-L_{i'j} = \tau_i^x - \tau_{i'}^x = C_{i,i'} \ ,$$

was eine starke Forderung darstellt. Bevor die Forderung interpretiert wird, soll deren Bedeutung für die Differenz $\tau_i^x - \tau_{i'}^x$ betrachtet werden. Ist $\tau_i^x > \tau_{i'}^x$, so gilt

$$\frac{1}{1+e^{\mu_j+\tau_i^x}} < \frac{1}{1+e^{\mu_j+\tau_{i'}^x}} \ ,$$

d.h.

$$F_i(j) < F_{i'}(j) \ , \quad j=1,\ldots,K-1.$$

Die Zufallsvariable Y_i ist stochastisch größer als $Y_{i'}$. Daraus folgt, daß die Verteilungsfunktionen sich nicht überschneiden dürfen. Aus der Anordnung der τ_i^x, i=1,...,R, erhält man eine Anordnung der Objekte nach der Lage der Verteilungsfunktionen auf der geordneten K-Punkte-Skala. Je größer τ_i^x ist, um so weiter rechts liegt die Verteilung von Y_i. D.h., das Objekt O_i wird meistens in die hohen Kategorien eingeordnet.

Gilt $\tau_i^x = 0$, i=1,...,R, so ist

$$L_{ij} = \mu_j, \ j=1,\ldots,K-1 \ ,$$

oder

$$L_{ij} = \log \frac{1-F(j)}{F(j)} \quad .$$

Alle Objekte haben dieselbe Urteilsverteilung, damit dieselbe Merkmalsausprägung.

Schließlich werde noch die Modellforderung

$$L_{ij} - L_{i'j} = \log \left(\frac{m_{i,j+1} + \ldots + m_{iK}}{m_{i1} + \ldots + m_{ij}} \Big/ \frac{m_{i',j+1} + \ldots + m_{i'K}}{m_{i'1} + \ldots + m_{i'j}} \right) = C_{i,i'}$$

für $j=1,\ldots,K-1$

interpretiert.

$C_{i,i'}$ kann als Assoziation zwischen den Zeilen i,i' interpretiert werden. Hierfür denke man sich die Antwortvariable dichotomisiert (Antwort $\leq j$, Antwort $> j$). Mit den Zeilen i und i' entsteht eine 2×2-Felder-Tafel. Es werde bezeichnet:

$$f_{11}^{i'} = m_{i'1} + \ldots + m_{i'j}, \quad f_{22}^{i} = m_{i,j+1} + \ldots + m_{iK},$$

$$f_{12}^{i'} = m_{i',j+1} + \ldots + m_{i'K}, \quad f_{21}^{i} = m_{i1} + \ldots + m_{ij} .$$

$$c_{i,i'} = \frac{f_{11}^{i'} f_{22}^{i}}{f_{12}^{i'} f_{21}^{i}}$$

ist das "odds ratio" der Vierfeldertafel.
Für jedes $j=1,\ldots,K-1$ erhält man eine solche Vierfeldertafel.

Das angegebene Logit-Modell setzt voraus, daß für zwei beliebige Zeilen i und i' die (K-1) Vierfeldertafel zu demselben

$$C_{i,i'} = \log c_{i,i'} = \log \left(\frac{f_{11}^{i'} f_{22}^{i}}{f_{12}^{i'} f_{21}^{i}} \right)$$

führen.

Mit den vorangehenden Betrachtungen kann gesagt werden, daß das spezielle Logit-Modell

$L_{ij} = \mu_j + \tau_i^x$, $i=1,\ldots,R$, $j=1,\ldots,K-1$

zur Skalierung von Objekten (bzw. Items) geeignet ist. Es stellt ein abgeleitetes Meßmodell dar. Bezeichnet man mit B die Menge der Objekte, mit $p_i = (p_{i1},\ldots,p_{iK})$ die Wahrscheinlichkeitsverteilung, welche durch die Beurteilung des Objektes O_i entsteht, so stellt $\mathfrak{B} = \langle B, (p_1,\ldots,p_R)\rangle$ ein abgeleitetes Meßsystem dar.

Für jede Zelle (i,j) der Kontingenztafel gibt es eine repräsentierende Relation $L_{ij} = \mu_j + \tau_i^x$, welche die Beziehung zwischen den Wahrscheinlichkeiten und den abgeleiteten Meßwerten τ_i^x herstellt. $\langle \mathfrak{B}, R, \tau\rangle$ ist eine abgeleitete Skala. Die repräsentierende Relation stellt eine psychologische Hypothese über das Verhalten der Rater in Bezug auf die Objekte dar. Diese Hypothese ist überprüfbar.

Das hier untersuchte Modell ergibt eine abgeleitete Ratio-Skala. Andere Modelle werden zu anderen abgeleiteten Skalen führen. Über das Skalenniveau der Fundamentalskala kann zunächst nichts ausgesagt werden.

Wenn man annimmt, daß Rater einen ordinalen Meßvorgang durchführen können, so bedeutet die Beurteilung der Objekte auf der geordneten K-Punkte-Skala ein fundamentales Messen. Von daher ist die Fundamentalskala des beurteilten Merkmals eine Ordinalskala.

Die τ_i^x, $i=1,\ldots,R$, sind Parameter der Urteilsverteilungen der zugehörigen Objekte, und zwar ordinale Parameter. Ein *ordinaler Parameter* kann dabei wie folgt definiert werden (Schneeweiß, 1967, S. 103):

"Ein Verteilungsparameter α heiße ordinaler Parameter, wenn zu je zwei Wahrscheinlichkeitsverteilungen F_1 und F_2, für die α den gleichen Wert annimmt ($\alpha[F_1] = \alpha[F_2]$), auch die durch irgendeine streng monoton steigende Funktion f transformierten Wahrscheinlichkeitsverteilungen F_{1f} und F_{2f} hinsichtlich dieses Parameters einander gleich sind ($\alpha[F_{1f}] = \alpha[F_{2f}]$)".

Ersetzt man α durch τ, so hat man das gewünschte Ergebnis. Da $F_{i'}(j) < F_i(j)$ für $\tau_{i'}^x < \tau_i^x$, gilt, handelt es sich um einen ordinalen Lageparameter. Die τ_i^x liegen als abgeleitete Meßwerte zwar auf einer Ratio-Skala, für die Fundamentalskala geben sie jedoch nur die Ordnung der Objekte an.

Das Vorgehen kann mit der Methode der gleicherscheinenden Intervalle verglichen werden. Die Ausgangsdaten, nämlich die von Ratern erstellten Häufigkeitsverteilungen für die Objekte, sind in beiden Fällen gleich. Wird ein Logit-Modell zur Skalierung verwendet, so müssen die Verteilungen gewissen Bedingungen genügen. Diese sind in der repräsentierenden Relation $L_{ij} = \mu_j + \tau_i^x$ wiedergegeben, welche empirisch überprüfbar ist. Bei der Methode der gleicherscheinenden Intervalle wird in den Häufigkeitsverteilungen der jeweilige Median zugeordnet. Diese repräsentierende Relation ist eine Definition, also empirisch nicht überprüfbar.

Die Anwendung eines Logit-Modells ist durch die Modellanpassung und die Schätzung der Parameter aufwendiger als z.B. die Methode der gleicherscheinenden Intervalle. Liegen jedoch speziell Items vor, die sich nach ihren Verteilungen anordnen lassen, so ist das hier vorgestellte Modell angemessen.

Konnte das Logit-Modell als probabilistische Version der Methode der gleicherscheinenden Intervalle aufgefaßt werden, so ist es auch möglich, es als probabilistische Version der Likert-Skalierung zu betrachten.
Wenn man die g Rater durch g Items ersetzt und statt R Objekten R Personen betrachtet, welche die Items auf einer geordneten K-Punkte-Skala beantworten, so werden mit dem Logit-Modell die Personen skaliert. Die Urteilsverteilungen der Personen müssen jetzt eine stochastische Ordnung aufweisen. Sowohl die Rater als auch die Items werden als homogen vorausgesetzt. Für die Rater bedeutet das, daß ihre Urteilsverteilungen bezüglich des Objektes O_i,

$i=1,\ldots,R$, als gleich vorausgesetzt werden. Die Rater sollen unabhängig sein und identische Urteilsverteilungen haben. Für die Items bestünde dieselbe Forderung.

Das betrachtete Logit-Modell gehört zu einer Klasse von Modellen, welche stochastische Ordnungen identifizieren. Diese Modelle wurden allgemein von McCullagh (1980) als Regressionsmodelle formuliert. Sie haben die allgemeine Form

$$\text{link}\{F_i(j)\} = \theta_j - \tau_i \ ,$$

wobei "link" eine monoton steigende Funktion bedeutet, welche das Intervall (0,1) auf $(-\infty,\infty)$ abbildet. Z.B. kann es sich um eine Logit-Transformation, um die Inverse der Verteilungsfunktion der Normalverteilung, die Inverse der Verteilungsfunktion der Cauchyverteilung oder die Funktion $\log(-\log(1-F_i(j)))$ handeln.

2.2. Äquivalenz zwischen probabilistischen Meßmodellen und Log-linearen Modellen

2.2.1. Das Rasch Modell

Es seien zwei Objektmengen gegeben. B_1 sei eine Menge von Personen, B_2 eine Menge von Aufgaben. Die Personen beantworten die Aufgaben auf einer dichotomen ordinal skalierten Antwortvariablen. Angenommen, es ist eine Kontingenztafel gegeben mit zwei Einflußfaktoren und einer dichotomen Antwortvariablen. Die Einflußfaktoren seien eine Personen- und eine Itemvariable mit C, j=1,...,C, bzw. R, i=1,...,R, Kategorien. Jede Person einer Stichprobe kann damit klassifiziert werden nach ihrem Wert auf der Personenvariablen und nach den gelösten bzw. nicht gelösten Aufgaben. Für die Kontingenztafel kann ein Log-lineares Modell vorgeschlagen werden. In dem vorliegenden Spezialfall von nur 2 Antwortkategorien hat ein Log-lineares Modell gleiche Struktur wie ein lineares logistisches Modell und dieses wiederum ist äquivalent einem Regressionsmodell (McCullagh, 1980).

Da im vorliegenden Fall nur die Effekte der Einflußfaktoren auf die Antwortvariable von Interesse sind, ist statt eines Log-linearen Modells ein Logit-Modell angemessener (Fienberg, 1977, S. 97). Damit wird die Antwortvariable als ordinale Variable behandelt. Ein solches Modell ist (Fienberg, 1977, Kap. 6):

$$L_{ij} = \log\left(\frac{m_{ij2}}{m_{ij1}}\right) = w + w_{1(i)} + w_{2(j)} + w_{12(ij)}$$

mit

$$\sum_i w_{1(i)} = \sum_j w_{2(j)} = \sum_i w_{12(ij)} = \sum_j w_{12(ij)} = 0 \ .$$

Für alle Zellen (i,j) wird das Verhältnis der Erwartungswerte der Besetzungszahlen der beiden Kategorien der Antwortvariablen gebildet. Dabei sei k=2 die positive Beantwortung des Items. Große Logits bedeuten damit für die An-

werte als für die Anzahl der negativ beantworteten Items.

$w_{1(i)}$, $i=1,\ldots,N$, sind die Effekte der Items, $w_{2(j)}$, $j=1,\ldots,C$, die Effekte der Personen und $w_{12(ij)}$ die Effekte der Item-Personen-Kombinationen auf das Antwortverhältnis.

In diesem saturierten Modell ist der Assoziationsterm $w_{12(ij)}$ für die Assoziation zwischen Items und Personen enthalten. Geht man davon aus, daß eine solche Assoziation gleich Null sein soll, so kann das folgende nichtsaturierte Modell angepaßt werden:

$$L_{ij} = \log\left(\frac{m_{ij2}}{m_{ij1}}\right) = w + w_{1(i)} + w_{2(j)}$$

mit

$$\sum_i w_{1(i)} = \sum_j w_{2(j)} = 0 \ .$$

Dieses Modell kann analog zu demjenigen im vorhergehenden Abschnitt 2.1.1. untersucht werden.

Durch Summation über i und j erhält man

$$\sum_{ij} L_{ij} = C\,N\,w \ ,$$

$$w = \frac{1}{C\,N} \sum_{ij} L_{ij} = \mu_{..} \ .$$

Summation über i ergibt die Personeneffekte

$$w_{2(j)} = \frac{1}{N} \sum_i L_{ij} - \mu_{..} \ ,$$

Summation über j die Itemeffekte

$$w_{1(i)} = \frac{1}{C} \sum_j L_{ij} - \mu_{..} \ .$$

Bildet man die Differenz der Logits für zwei beliebige Zeilen i und i', so ergibt sich

$$L_{ij} - L_{i'j} = w_{1(i)} - w_{1(i')} = \log\left(\frac{m_{ij2}}{m_{ij1}} \frac{m_{i'j1}}{m_{i'j2}}\right) .$$

Das Modell fordert, daß diese Differenzen für alle j, also für alle Kategorien der Personenvariablen, denselben Wert annehmen. Ist $w_{1(i)} > w_{1(i')}$, so muß dieses für alle Kategorien j gelten. D.h., für alle j muß das Item i im Erwartungswert häufiger positiv beantwortet werden als das Item i'. Diese Modelleigenschaft kann zur Itemanalyse verwendet werden.

Entsprechend gilt für die Differenz der Logits für zwei beliebige Kategorien j und j':

$$L_{ij} - L_{ij'} = w_{2(j)} - w_{2(j')} = \log \left(\frac{m_{ij2}}{m_{ij1}} \frac{m_{ij'1}}{m_{ij'2}} \right) .$$

Die Differenz muß jetzt für alle Items i denselben Wert annehmen. Ist $w_{2(j)} > w_{2(j')}$, so müssen die Personen, die zur Kategorie j gehören, im Erwartungswert alle Items i häufiger positiv beantworten als die Personen, die zur Kategorie j' gehören.

Diese starken Forderungen sind bei Irtel und Schmalhofer (1981/82) abgemildert. Sie fordern nur Ordnungen zwischen den Zellwahrscheinlichkeiten p_{ijk}. Damit können sie Items und Personen getrennt voneinander auf Ordinalskalen anordnen. Durch Anwendung des Logit-Modells werden Items und Personen gemeinsam bezüglich des Merkmals "Effekt auf das Antwortverhältnis" angeordnet. Der Itemeffekt ist um so größer, je häufiger das Item positiv beantwortet wurde. Damit erhält eine große "Itemleichtigkeit" einen hohen Effekt. Da $\sum_i w_{1(i)} = 0$ ist, erhält man durch Umkehren der Vorzeichen der geschätzten Effekte die "Itemschwierigkeit".

Das Logit-Modell ist ein abgeleitetes Meßmodell mit 2 Objektmengen B_1 und B_2. Die Wahrscheinlichkeiten p_{ij} sind auf $B_1 \times B_2$ definiert.

$$L_{ij} = \log \left(\frac{m_{ij2}}{m_{ij1}} \right) = \log \left(\frac{p_{ij2}}{p_{ij1}} \right) = w + w_{1(i)} + w_{2(j)}$$

ist die repräsentierende Relation für die Zelle (i,j). $w_{1(i)}$, i=1,...,R, und $w_{2(j)}$, j=1,...,C, sind abgeleitete Meßwerte, die auf einer gemeinsamen, abgeleiteten Ratio-

Bei der Anwendung des Modells wird es nicht immer möglich sein, die Personen einer Stichprobe nach den Werten auf einer Personenvariablen zu klassifizieren. Die Itemvariable kann als zunächst nominaler Einflußfaktor mit R Ausprägungen (Aufgaben) aufgefaßt werden. Als Personenvariable wird üblicherweise die "Anzahl der gelösten Aufgaben" gewählt. Bei R Aufgaben kann die Anzahl der gelösten Aufgaben von 0 bis R variieren. Das ergäbe (R+1) Scoregruppen. Hiervon werden diejenigen Spalten der Kontingenztafel weggestrichen, welche aus leeren Zellen bestehen und diejenigen Aufgaben, welche von allen Personen gleich beantwortet werden (Mellenbergh und Vijn, 1981).

Das Logit-Modell würde hier als Personeneffekt den Effekt der Anzahl der gelösten Aufgaben auf das Antwortverhältnis (gelöst zu nicht gelöst) schätzen. Das erscheint nur dann sinnvoll, wenn die Zahl der gelösten Aufgaben einer Person etwas mit deren Fähigkeit zu tun hat. Das kann mit der Forderung des Rasch Modells, daß die Gesamtzahl der gelösten Aufgaben eine erschöpfende Statistik für den Personenparameter ist, verglichen werden.
Um hier überhaupt ein Schema zu erhalten, welches als Kontingenztafel aufgefaßt werden kann und damit die geschätzten Parameter für den Einflußfaktor "Zahl der gelösten Aufgaben" sinnvoll interpretiert werden können, muß eine latente Variable "Fähigkeit" angenommen werden. Notwendig ist weiterhin die Forderung, daß die Summe der gelösten Aufgaben eine erschöpfende Statistik für die Fähigkeit ist. Derartige Annahmen sind nicht üblich bei Log-linearen Modellen. Mellenbergh und Vijn (1981) zeigten, daß für eine Kontingenztafel, die so gebildet wird, wie es oben beschrieben wurde, das angegebene Logit-Modell dem Rasch Modell entspricht.
Die Wahrscheinlichkeit p_{ij}, daß eine Person j ein Item i positiv beantwortet, ist nach dem Rasch Modell

$$p_{ij} = \frac{\exp(a_j - d_i)}{1+\exp(a_j - d_i)} \quad .$$

Dabei ist a_j die Personenfähigkeit, d_i die Itemschwierig-

keit. Daraus folgt

$$\frac{p_{ij}}{1-p_{ij}} = e^{a_j - d_i}$$

und

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = a_j - d_i \ .$$

Das angegebene Logit-Modell lautet

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \log\left(\frac{p_{ij2}}{p_{ij1}}\right) = \log\left(\frac{m_{ij2}}{m_{ij1}}\right) = w + w_{1(i)} + w_{2(j)}$$

mit

$$\sum_i w_{1(i)} = \sum_j w_{2(j)} = 0 \ .$$

Der Fähigkeitsparameter a_j des Rasch Modells entspricht dem Personeneffekt des Logit-Modells, der Itemparameter d_i des Rasch Modells entspricht dem Aufgabeneffekt des Logit-Modells. Das Logit-Modell enthält zusätzlich die Konstante w und die Nebenbedingungen.
Während bisher nur die Modellgleichung des Rasch Modells betrachtet wurde, soll es jetzt als ein probabilistisches Testmodell gesehen werden. Für diese Modelle wird die Existenz einer latenten, nicht direkt beobachtbaren Dimension ξ postuliert. Jede dichotom zu beantwortende Aufgabe i ist mit der latenten Dimension ξ durch eine Funktion $f_i(\xi)$ verknüpft. Diese hat im Falle des Rasch Modells für die Person ν und die Aufgabe i die Form

$$f_i(\xi) = \frac{e^{(\xi_\nu - \sigma_i)}}{1+e^{(\xi_\nu - \sigma_i)}} \ .$$

Dabei sind ξ_ν und σ_i Parameter bezüglich der latenten Dimension (Fischer, 1974).

Der Unterschied zur Kontingenztafelanalyse wird deutlich. Bei der Analyse von Kontingenztafeln werden üblicherweise keine latenten Variablen vorausgesetzt. Die Parameter $w_{1(i)}$ und

Aus diesem Grunde unterscheidet sich auch die Interpretation der Wahrscheinlichkeiten p_{ij1} und p_{ij2} von derjenigen beim Rasch Modell. Auch die Art der Parameterschätzung ist unterschiedlich, obwohl die Schätzmethoden, die bei Loglinearen Modellen verwendet werden, auch beim Rasch Modell verwendet werden können.

Obwohl die Modellgleichungen des Rasch Modells und des Logit-Modells gleich sind, kann lediglich von einer formalen Äquivalenz der Modelle gesprochen werden, da beim Rasch Modell als einem probabilistischen Testmodell latente Variablen angenommen werden und damit die Parameter anders interpretiert werden. Das Logit-Modell kann als eigenständiges abgeleitetes Meßmodell angesehen werden.

2.2.2. Das BTL-Modell

K Objekte sollen von g Urteilern bezüglich ihrer Merkmalsausprägung miteinander verglichen werden. Die Objekte werden den Urteilern in $\binom{K}{2}$ Paaren dargeboten, und jeder Urteiler gibt seine Präferenz an. Die Entscheidung "keine Präferenz" sei zunächst nicht erlaubt. Es wird angenommen, daß die Rater unabhängig voneinander urteilen und daß ein Rater die einzelnen Paare unabhängig voneinander beurteilt.

Das Bradley-Terry-Modell postuliert für die Wahrscheinlichkeit, daß Objekt O_i dem Objekt O_j vorgezogen wird

$$P(O_i \succ O_j) = p_{ij} = \frac{v_i}{v_i + v_j} \quad , \; i \neq j, \; i,j = 1,\ldots,K,$$

$$p_{ij} + p_{ji} = 1, \; p_{ii} = \frac{1}{2} \; .$$

Im folgenden sollen verschiedene Möglichkeiten, Log-lineare Modelle auf die Paarvergleichsdaten anzuwenden, betrachtet werden.

Betrachtet man die beiden Einflußfaktoren "Beurteiltes Paar" mit den Ausprägungen $i=1,\ldots,\binom{K}{2}$, "1. Objekt im Paar" mit den Ausprägungen $O_1,\ldots,O_K$ und eine dichotome Antwortvariable mit den Ausprägungen (nicht gelöst, gelöst) oder die

beiden Einflußfaktoren "1. Objekt im Paar", "2. Objekt im Paar" und die dichotome Antwortvariable, so können die Paarvergleichsdaten in einer Kontingenztafel dargestellt werden. Hier wird die Stellung des Objektes im Paar berücksichtigt, was beim BTL-Modell nicht der Fall ist. Auf diese Weise könnte überprüft werden, ob der Platz im dargebotenen Paar einen Einfluß auf das Urteil hat.

Die vorgeschlagenen Kontingenztafeln enthalten viele Nullen, die als "strukturelle Nullen" bezeichnet werden, da sie Folgen aus dem Modell sind. Nach Bishop et al. (1975, S. 210) kann an eine solche Tafel ein Quasi-Log-lineares Modell angepaßt werden. Das bedeutet, daß nur die Menge S derjenigen Zellen der dreidimensionalen Tafel betrachtet wird, die keine strukturellen Nullen enthalten. Für diese Zellen ist m_{ijk} der Erwartungswert für die Zelle (i,j,k), es ist $m_{ijk} = 0$ für $(i,j,k) \notin S$. Für die Zellen aus S wird ein allgemeines Log-lineares Modell formuliert

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)}$$

für $(i,j,k) \in S$ mit

$$\sum_i \delta_i^{(23)} u_{1(i)} = \sum_i \delta_{ij}^{(3)} u_{12(ij)} = \sum_i \delta_{ik}^{(2)} u_{13(ik)} = \sum_i \delta_{ijk} u_{123(ijk)} = 0 \ ,$$

usw.

Dabei ist

$$\delta_{ijk} = \begin{cases} 1 & \text{für } (i,j,k) \in S \\ 0 & \text{sonst} \end{cases} ,$$

$$\delta_{ij}^{(3)} = \begin{cases} 1 & \text{wenn } \delta_{ijk} = 1 \text{ für ein } k \\ 0 & \text{sonst} \end{cases} ,$$

$$\delta_i^{(23)} = \begin{cases} 1 & \text{wenn } \delta_{ijk} = 1 \text{ für ein } (j,k) \\ 0 & \text{sonst} \end{cases} ,$$

ähnlich für $\delta_{ik}^{(2)}$, $\delta_{jk}^{(1)}$, $\delta_{j}^{(13)}$, $\delta_{k}^{(12)}$.

Der hochgestellte Index zeigt an, ob i, j oder k bei den tiefgestellten Indizes fehlen. Z.B. fehlen bei $\delta_{j}^{(13)}$ das i und das k, also der 1. und der 3. Index.
Die Bedingungen sind dieselben wie für ein übliches Loglineares Modell, gelten hier aber nur für die Zellen aus S.

Da nur der Einfluß auf die Antwortvariable von Interesse ist, erscheint es angemessener, ein Logit-Modell statt eines Loglinearen Modells zu betrachten. Ein Logit-Modell erhält man z.B. durch Bilden der Differenz $\log m_{ij2} - \log m_{ij1}$ und Zusammenfassen und Neubenennung der Summanden:

$$L_{ij} = \log \left(\frac{m_{ij2}}{m_{ij1}} \right) = w + w_{1(i)} + w_{2(j)} + w_{12(ij)}$$

mit den entsprechend formulierten Restriktionen für die Zellen aus S.
Das nichtsaturierte Modell lautet

$$L_{ij} = \log \left(\frac{m_{ij2}}{m_{ij1}} \right) = w + w_{1(i)} + w_{2(j)} \ .$$

Dieses Modell wird mit der Modellgleichung des BTL-Modells verglichen. Es ist

$$\frac{p_{ij}}{p_{ji}} = \frac{v_i}{v_j} \ ,$$

$$\log \left(\frac{p_{ij}}{p_{ji}} \right) = \log \left(\frac{m_{ij}}{m_{ji}} \right) = \log v_i - \log v_j \ .$$

Beim BTL-Modell fehlt die Konstante, sowie die Restriktionen des Logit-Modells. Die Interpretation der Wahrscheinlichkeiten ist unterschiedlich und damit die Interpretation der Parameter. Das ist jedoch zu erwarten, da das BTL-Modell die Stellung im Paar nicht berücksichtigt.
Während beim Logit-Modell $w_{1(i)}$ den Einfluß des Faktors "Beurteiltes Paar" auf das Antwortverhältnis und $w_{2(j)}$ den Einfluß des Faktors "1. Objekt" darstellt, beziehen sich sowohl v_i wie v_j auf die Objekte. Die Ähnlichkeit der Mo-

dellgleichungen ist größer, wenn $w_{1(i)}$ den Einfluß des 1. Objektes im Paar, $w_{2(j)}$ den Einfluß des 2. Objektes im Paar darstellt. Hat der Platz im Paar keinen Einfluß auf die Antwort, so beziehen sich beide Parameter auf die Objekte, wie im BTL-Modell.

Von einer Äquivalenz zwischen dem Logit-Modell und dem BTL-Modell kann schon wegen der unterschiedlichen Voraussetzungen nicht gesprochen werden. Das Logit-Modell kann aber unabhängig vom BTL-Modell bezüglich seiner Leistungsfähigkeit in diesem Rahmen untersucht werden.

Fienberg und Larntz (1976) geben zwei mögliche Darstellungen der Paarvergleichsdaten:

Tab. 2.3.: Darstellung der Paarvergleichsdaten für K=4

a)

nicht bevorzugt	bevorzugt O_1	O_2	O_3	O_4
O_1	-	n_{12}	n_{13}	n_{14}
O_2	n_{21}	-	n_{23}	n_{24}
O_3	n_{31}	n_{32}	-	n_{34}
O_4	n_{41}	n_{42}	n_{43}	-

b)

bevorzugte Objekte	Paar-Nr. 1	2	3	4	5	6
O_1	n_{12}	n_{13}	n_{14}	-	-	-
O_2	n_{21}	-	-	n_{23}	n_{24}	-
O_3	-	n_{31}	-	n_{32}	-	n_{34}
O_4	-	-	n_{41}	-	n_{42}	n_{43}

Beide Darstellungen sind Kontingenztafeln mit den Restriktionen $n_{ij}+n_{ji} = g$. Beide Tafeln enthalten strukturelle Nullen. Für die Menge S der Zellen, welche keine strukturellen Nullen enthalten, kann jeweils ein quasi-Log-lineares Modell angepaßt werden.
Die Tafel a) ist eine $K \times K$-Tafel, die Tafel b) vom Format $K \times \binom{K}{2}$.

Es wird zunächst die quadratische Tafel betrachtet. Es ist

$$\frac{m_{ij}}{m_{ji}} \frac{m_{jk}}{m_{kj}} \frac{m_{ki}}{m_{ik}} = \frac{v_i}{v_j} \frac{v_j}{v_k} \frac{v_k}{v_i} = 1 \quad .$$

Dieses stellt die Definition der Quasisymmetrie für quadratische Tafeln dar (Bishop et al., 1975, S. 287). Damit entspricht in diesem Fall die Modellgleichung des BTL-Modells einem Quasisymmetrie-Modell, welches additiv oder multiplikativ formuliert wird.

Nach Fienberg und Larntz (1976) ist die iterative Methode zum Schätzen der Parameter nach Darroch und Ratcliff (1972) äquivalent zu derjenigen, welche Bradley und Terry benutzten.

Für die Tafel b) empfehlen Fienberg und Larntz (1976) ein Modell der Quasiunabhängigkeit (Bishop et al., 1975, S.178). Der Begriff der "quasi independence" wurde formal von Goodman (1968) für 2×2-Tafeln eingeführt. Quasiunabhängigkeit bedeutet ein Unabhängigkeitsmodell für die Zellen ohne strukturelle Nullen. Es lautet

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} \quad .$$

Das Modell kann auch für die quadratische Tafel verwendet werden. Für quadratische Tafeln umfaßt die Quasiunabhängigkeit die Quasisymmetrie (Bishop et al., 1975, S. 288).

Das BTL-Modell für Paarvergleiche stellt ein abgeleitetes Meßmodell dar (Suppes und Zinnes, 1963). Die Menge der Ob-

jekte, welche paarweise miteinander verglichen werden sollen, werde mit B bezeichnet. Für alle Paare O_i, O_j ist p_{ij} die Wahrscheinlichkeit, mit welcher O_i dem O_j vorgezogen wird. Die Wahrscheinlichkeiten sind auf $B \times B$ definiert. $\mathfrak{B} = \langle B, p \rangle$ ist ein abgeleitetes Meßsystem.

Je nach Art der repräsentierenden Relation erhält man abgeleitete Skalen unterschiedlichen Niveaus. Suppes und Zinnes (1963) geben repräsentierende Relationen für abgeleitete Skalen auf Ratio-, Log-Intervall-, Intervall- und Differenzskalenniveau an.

Die hier betrachtete repräsentierende Relation lautet

$$p_{ij} = \frac{v_i}{v_i + v_j} \quad .$$

Sie bedeutet eine Hypothese über die Urteile der Rater, die empirisch überprüfbar ist. Die v_i, $i=1,\ldots,K$, stellen abgeleitete Meßwerte dar. Die abgeleitete Skala hat Ratio-Skalenniveau. Die abgeleitete Messung sagt nichts über das Skalenniveau der Fundamentalskala aus. Da das Modell jedoch eine Ordnung der Objekte bezüglich ihrer Merkmalsausprägungen voraussetzt, und die Rater durch die Beurteilung eine Messung auf Ordinalskalenniveau vornehmen, wird für die Fundamentalskala ein ordinales Skalenniveau angenommen.

Auch die betrachteten Log-linearen Modelle stellen abgeleitete Meßmodelle dar, die entsprechende abgeleitete Skalen erstellen. Für die Fundamentalskala wird mit demselben oben aufgeführten Argument ein ordinales Skalenniveau angenommen.

Bezüglich der Äquivalenz von BTL-Modell und Log-linearen Modellen kann man zunächst sagen, daß sowohl das BTL-Modell wie auch die Log-linearen Modelle abgeleitete Skalen für die Objekte erstellen. In beiden Fällen kann für die Fundamentalskala nur ordinales Skalenniveau angenommen werden.

Die Modelle für die dreidimensionalen Kontingenztafeln stellen eine Verallgemeinerung des BTL-Modells dar, da sie den [illegible] Objekte im Raum berücksichtigen.

Für das Quasisymmetrie-Modell und für das Quasiunabhängigkeitsmodell liegt formal eine Äquivalenz mit dem BTL-Modell vor.
Da bei den Log-linearen Modellen keine latenten Variablen angenommen werden, ist die Interpretation der Parameter in jedem Fall unterschiedlich zu der des BTL-Modells. Statt von einer Äquivalenz von Modellen zu sprechen, sollte man eher von einer Anwendung von Kontingenztafelmethoden auf die Paarvergleichsdaten sprechen.
Die Darstellung eines Skalierungsmodells als Log-lineares Modell kann von Vorteil sein. So gaben Fienberg und Larntz (1976) eine Verallgemeinerung auf multiple Paarvergleiche an.
Fienberg (1979) gab eine Log-lineare Darstellung von Paarvergleichsmodellen mit Bindungen, Beaver (1983) eine Log-lineare Darstellung für multivariate Paarvergleiche mit Bindungen.

2.4. Zusammenfassung

In den vorangehenden Abschnitten wurden einige Log-lineare Modelle als Skalierungsmodelle bzw. Meßmodelle aufgefaßt. Dabei wurde zunächst eine Einführung in die Theorie der Log-linearen Modelle gegeben. Auf eine Darstellung der historischen Entwicklung von Skalierungsmodellen wurde verzichtet. Diese wurde als bekannt vorausgesetzt.
Unter einem "Log-linearen Modell" wird eine Methode zur Analyse von Kontingenztafeln verstanden. Annahmen bezüglich des Skalenniveaus der Variablen in der Kontingenztafel oder bezüglich latenter Variablen, werden nicht gemacht.

Bei probabilistischen Skalierungs- bzw. Meßmodellen werden in der Regel latente Variablen vorausgesetzt. Von daher kann i.a. von vornherein keine Äquivalenz zwischen Log-linearen Modellen und probabilistischen Meßmodellen bestehen. Eine Äquivalenz kann höchstens bezüglich der formalen Aspekte der Modelle vorhanden sein.
Unabhängig von einer vorhandenen Äquivalenz kann alternativ zu bekannten Skalierungsmodellen ein entsprechendes Log-lineares Modell als abgeleitetes Meßmodell zur Skalierung verwendet werden. Voraussetzung ist, daß die Daten sich in einer Kontingenztafel darstellen lassen. Da in beiden Fällen für die Fundamentalskala eine Ordinalskala angenommen wird, ist anzunehmen, daß für die Objekte dieselbe Ordnung entsteht, unabhängig davon, ob ein probabilistisches Meßmodell oder ein Log-lineares Modell verwendet wird. Die Interpretation der Parameter bezieht sich beim Log-linearen Modell auf das beobachtete Merkmal, beim probabilistischen Modell auf die latente Variable.
Die Verwendung von Methoden der Kontingenztafelanalyse in der Skalierung kann von Vorteil sein. Selbst wenn keine Äquivalenz zu bekannten Verfahren vorliegt, können sich neue Möglichkeiten der abgeleiteten Messung ergeben, eventuell können Schätzverfahren übernommen werden oder es werden durch die Darstellung bekannter Verfahren als Log-lineares Modell Verallgemeinerungen möglich.

C. Methodisches Resumée

Die beiden hier betrachteten Methoden (bzw. Methodenkomplexe), das ordinale Testmodell (Schulman, 1975) und Log-lineare Modelle als Skalierungsmodelle, sind von unterschiedlicher Natur. Meßtheoretisch gesehen stellt das ordinale Testmodell ein abgeleitetes Meßverfahren mit empirisch nicht überprüfbarer repräsentierender Relation dar. Hierin unterscheidet es sich nicht von der Klassischen Testtheorie. Während die Scores in der Klassischen Testtheorie jedoch so behandelt werden, als handele es sich um fundamentale Messungen auf Intervallskalenniveau, handelt es sich beim ordinalen Testmodell um Rangzahlen. Es wurde angenommen, daß die Rangzahlen durch Beurteilung der zu skalierenden Objekte mit Hilfe von Ratern entstehen. Diese nehmen die fundamentale Messung vor. Das ordinale Testmodell ist bezüglich der adäquaten Behandlung der Daten der Klassischen Testtheorie vorzuziehen. Es stellt eine Übertragung der Konzepte der Klassischen Testtheorie auf Rangzahlen dar. Die Konzepte der Klassischen Testtheorie, wie z.B. lokale Unkorreliertheit, lineare experimentelle Unabhängigkeit, experimentelle Unabhängigkeit, Reliabilität, Paralleltests und Kriteriumsvalidität beziehen sich dadurch auf das Urteilsverhalten der Rater. So stellt z.B. die Reliabilität die Meßgenauigkeit der Rater dar, also die Urteilerübereinstimmung. Das ordinale Testmodell kann damit auch als Theorie des Raterverhaltens aufgefaßt werden. Als Urteilsmodell ist das Modell jedoch recht einfach. Z.B. berücksichtigt das Modell $R_i = t_i + E_i$ für das Urteil des Raters bezüglich des Individuums i nicht die Herkunft des Meßfehlers. Dafür können beispielsweise die Rater verantwortlich sein und/oder das beurteilte Individuum oder zusätzlich eine Wechselwirkung von Rater und Individuum.

Daß es möglich ist, die Konzepte der Klassischen Testtheorie auf Rangzahlen zu übertragen, liegt an der Definition des wahren Wertes t_i eines Individuums i als Erwartungswert der Rangzahlen. Die Differenz zwischen beobachtetem Wert R_i und dem wahren Wert t_i wird als Meßfehler E_i definiert. Die Besonderheiten dieses Modells ergeben sich beim Schätzen von Populationsparametern mit

Hilfe einer Zufallsstichprobe.

Will man bei bekannter Population z.B. die wahren Werte der beurteilten Objekte berechnen, so muß die Wahrscheinlichkeitsverteilung der Rangreihen bekannt sein. Schulman (1979) gibt ein einfaches Modell für eine solche Verteilung an, die keine Gleichverteilung der möglichen Rangpermutationen ist (siehe auch Tallis und Dansie, 1983). Dieser Aspekt wurde in der vorliegenden Arbeit nicht berücksichtigt. Die Rangreihen stellen in der Skalierung die Daten dar, auf welche das Skalierungsverfahren angewendet wird. Eine Wahrscheinlichkeitsverteilung für die Rangreihen stellt eine Urteilsverteilung der Rater dar und ist somit auch ein Modell für das Raterverhalten.

Während das ordinale Testmodell für Merkmale mit vielen Ausprägungen geeignet ist, daß die Bildung einer Rangreihe mit möglichst wenig Bindungen möglich ist, stellt das betrachtete Logit-Modell ein Verfahren für kategorielle Daten dar. In seiner Funktion als Skalierungsmodell stellt es ein abgeleitetes Meßmodell mit empirisch überprüfbarer repräsentierender Relation dar. Mit seiner Hilfe wird eine Ordnung der zu skalierenden Objekte erstellt. Das Modell kann als stochastische Version der Methode der gleicherscheinenden Intervalle gesehen werden. Während diese Methode immer angewandt werden kann, ist das Logit-Modell nur dann zur Skalierung verwendbar, wenn die Daten die Modellforderungen erfüllen. Die Verwendung Log-linearer Modelle in der Skalierung bedeutet, Modelle, die zur Analyse von Kontingenztafeln angewandt werden, jetzt im Rahmen der Psychometrie zu verwenden. Voraussetzung für die Anwendung ist dabei, daß die Daten sich in Form einer Kontingenztafel darstellen lassen.
Log-lineare Modelle machen keine Annahmen bezüglich des Skalenniveaus der Variablen in der Kontingenztafel oder bezüglich latenter Variablen. Bei probabilistischen Skalierungsmodellen werden i.a. latente Variablen angenommen. Eine Äquivalenz zwischen probabilistischen Meßmodellen und Log-linearen Modellen besteht von daher nicht.

Da für die betrachteten Methoden Rater von großer Bedeutung sind, seien abschließend hierzu einige Bemerkungen gemacht. In beiden

Rater gefordert. Hierunter wurden Rater verstanden, welche nach derselben Wahrscheinlichkeitsverteilung unabhängig voneinander urteilen. Beim ordinalen Testmodell wurden nur gleiche Erwartungswerte und Varianzen gefordert.

In der praktischen Anwendung kann die Homogenität der Rater nicht überprüft werden. Größere Unterschiede zwischen den Rangreihen können unterschiedliche Ursachen haben. So kann es sich z.B. um schlecht unterscheidbare Objekte, aber um kompetente Rater handeln oder um gut unterscheidbare Objekte und inkompetente Rater. Auch inkompetente Rater können homogen sein. Sie können z.B. alle nach derselben Wahrscheinlichkeitsverteilung urteilen, etwa einer Gleichverteilung für die möglichen Rangreihen. Eine ideale Situation läge vor, wenn die Rater homogen und kompetent und die Objekte gut unterscheidbar wären.
Die statistischen Verfahren im Abschnitt I.3.3. arbeiten mit der Nullhypothese der Gleichwahrscheinlichkeit der möglichen Rangreihen. Wenn man darunter die Inhomogenität der Rater verstehen will, so muß vorausgesetzt werden, daß diese kompetent und die Objekte gut unterscheidbar sind. Durch Ablehnung der Nullhypothese können eventuell homogene Untergruppen von Ratern nachgewiesen werden. Die Inhomogenität der Rater kann verschiedene Ursachen haben, z.B. unterschiedliche Einstellungen, unterschiedliches Vorwissen der Rater oder auch Mehrdimensionalität des Merkmals. Lassen die Rater in das Urteil, welches objektiv sein soll, z.B. zuviel von ihren Einstellungen oder Präferenzen einfließen, so erhält man eher Präferenzurteile als objektive Urteile. Das Schätzen der wahren Werte beim ordinalen Testmodell kann dann fragwürdig werden, da die Rater möglicherweise nicht dieselben Merkmale beurteilen, also unterschiedliche Urteilskriterien haben. Weiterhin ist zu vermuten, daß die Rater nur in gewissen Bereichen gut differenzieren können, daß das Urteil also von unterschiedlicher Güte ist. Um mit einer verläßlichen Beurteilung rechnen zu können, sollten ein Ratertraining und eine Rateranalyse durchgeführt werden.

D. Literaturverzeichnis

Agresti, A.: A survey of strategies for modeling cross-classifications having ordinal variables. Journal of the American Statistical Association 78, 1983, 184-198.

Agresti, A., Kezouh, A.: Association models for multi-dimensional cross-classification for ordinal variables. Communication in Statistics, Theory and Methods 12, 1983, 1261-1276.

Anderson, J.A., Philips, P.R.: Regression, discrimination and measurement models for ordered categorical variables. Applied Statistics 30, 1981, 22-31.

Anderson, R.L.: Use of contingency tables in the analysis of consumer preference studies. Biometrics 15, 1959, 582-590.

Andrich, D.: A model for contingency tables having an ordered response classification. Biometrics 35, 1979, 403-415.

Bartlett, M.S.: Contingency table interactions. Supplement of the Journal of the Royal Statistical Society 2, 1935, 248-252.

Beaver, R.J.: Log-linear models for multivariate paired comparison experiments with ties. Journal of Statistical Planning and Inference 7, 1983, 209-218.

Birch, M.W.: Maximum likelihood in three-way contingency tables. Journal of the Royal Statistical Society B 25, 1963, 220-233.

Bishop, Y.M.M., Fienberg, S.E., Holland, P.W.: Discrete multivariate analysis: theory and practice. Cambridge: MIT Press, 1975.

Brogden, H.E.: The Rasch model, the law of comparative judgement and additive conjoint measurement. Psychometrika 42, 1977, 631-634.

Campbell, N.R.: Physics: The elements. Cambridge: University Press 1920. Neu gedruckt unter dem Titel "Foundations of science", New York: Dover, 1957.

Clayton, D.G.: Some odds ratio statistics for the analysis of ordered categorical data. Biometrika 61, 1974, 525-531.

Coombs, C.H.: Psychological scaling without a unit of measurement. Psychological Review 57, 1950, 145-158.

Cronbach, L.J.: Coefficient alpha and the internal structure of tests. Psychometrika 16, 1951, 297-334.

Darroch, J.N., Ratcliff, D.: Generalized iterative scaling for log-linear models. Annals of Mathematical Statistics 43, 1972, 1470-1480.

Davison, M.L.: A psychological scaling model for testing order hypotheses. British Journal of Mathematical and Statistical Psychology 33, 1980, 123-141.

Fahrmeir, L., Hamerle, A.: Multivariate statistische Verfahren. Berlin: de Gruyter, 1984.

Fienberg, S.E.: The analysis of cross-classified categorical data. Cambridge: MIT Press, 1977.

Fienberg, S.E.: Loglinear representation for paired comparison models with ties and within-pair order effects. Biometrics 35, 1979, 479-481.

Fienberg, S.E., Larntz, K.: Loglinear representation for paired and multiple comparison models. Biometrika 63, 1976, 245-254.

Fischer, G.: Einführung in die Theorie psychologischer Tests. Bern: Huber, 1974.

Fischer-Cornelssen, K.A., Abt, K.: Videotape recording in psychiatry and psychopharmacology. Acta Psychiatrica Scandinavica 61, 1980, 228-238.

Fishburn, P.C.: Utility theory for decision making. New York: Wiley, 1970.

Friedman, M.: The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the American Statistical Association 32, 1937, 675-701.

Gigerenzer, G.: Messung und Modellbildung in der Psychologie. München: Ernst Reinhardt GmbH & Co., Verlag, 1981.

Goodman, L.A.: The analysis of cross-classified data: independence, quasi-independence, and interactions in contingency tables with or without missing entries. Journal of the American Statistical Association 63, 1968, 1091-1131.

Goodman, L.A.: A new model for scaling response patterns: An application of the quasi-independence concept. Journal of the American Statistical Association 70, 1975, 755-768.

Goodman, L.A.: Simple models for the analysis of association in cross-classifications having ordered categories. Journal of the American Statistical Association 74, 1979, 537-552.

Goodman, L.A.: The analysis of dependence in cross-classification having ordered categories using log-linear models for frequencies and log-linear models for odds. Biometrics 39, 1983, 149-160.

Haberman, S.J.: The analysis of frequency data. Chicago: University of Chicago Press, 1974.

Helmholtz, H. v.: Zählen und Messen, erkenntnistheoretisch betrachtet. In: Philosophische Aufsätze, Eduard Zeller zu seinem fünfzigjährigen Doctor-Jubiläum gewidmet. Leipzig: Fues' Verlag, 1887, 15-52.

Huber, H.P.: Psychometrische Einzelfalldiagnostik. Weinheim: Beltz, 1973.

Irtel, H., Schmalhofer, F.: Psychodiagnostik auf Ordinalskalenniveau: Meßtheoretische Grundlagen, Modelltests und Parameterschätzung. Archiv für Psychologie 134, 1981/82, 197-218.

Johnson, H.M.: Pseudo-mathematics in the mental and social sciences. American Journal of Psychology 48, 1936, 342-351.

Kannemann, K.: An incidence test for k related samples. Biometrische Zeitschrift 18, 1976, 3-11.

Kelderman, H.: Loglinear Rasch model tests. Psychometrika 49, 1984, 223-245.

Kendall, M.G.: Rank correlation methods. London: Charles Griffin & Company Limited, 1962.

Koch, G.G., Freeman, J.L., Lehnen, R.G.: A general methodology for the analysis of ranked policy preference data. International Statistical Review 44, 1976, 1-28.

Krantz, D.H., Luce, R.D., Suppes, P., Tversky, A.: Foundations of measurement, Vol. 1, New York: Academic Press, 1971.

Lehmacher, W., Wall, K.-D.: A new nonparametric approach to the comparison of K independent samples of response curves. Biometrical Journal 20, 1978, 261-273.

Lienert, G.A.: Verteilungsfreie Methoden in der Biostatistik. Band I. 2. Auflage. Meisenheim: Hain, 1973.

Lienert, G.A.: Verteilungsfreie Methoden in der Biostatistik. Band II. Meisenheim: Hain, 1978.

Likert, R.: A technique for the measurement of attitudes. Archives of Psychology 140, 1932, 1-55.

McCullagh, P.: Regression models for ordinal data. Journal of the Royal Statistical Society B 42, 1980, 109-142.

McGregor, D.: Scientific measurement and psychology. Psychological Review 42, 1935, 246-266.

Mellenbergh, G.J., Vijn, P.: The Rasch model as a loglinear model. Applied Psychological Measurement 5, 1981, 369-376.

Orth, B.: Einführung in die Theorie des Messens. Stuttgart: Kohlhammer, Standards Psychologie, 1974.

Pearson, K.: Mathematical contributions to the theory of evolution in the inheritance of characters not capable of exact quantitative measurement, VIII. Philosophical Transactions of the Royal Society of London, Series A 195, 1900, 79-150.

Perline, R., Wright, B.D., Wainer, H.: The Rasch model as additive conjoint measurement. Applied Psychological Measurement 3, 1979, 237-255.

Pfanzagl, J.: Theory of measurement. Würzburg: Physica Verlag, 1968.

Plackett, R.L.: The analysis of categorical data. London: Griffin & Company Limited, 1974.

Roy, S.N., Kastenbaum, M.A.: On the hypothesis of no "interaction" in a multi-way contingency table. Annals of Mathematical Statistics 27, 1956, 749-757.

Scates, D.E.: The essential conditions of measurement. Psychometrika 2, 1937, 27-34, a.

Scates, D.E.: How science measures. Journal of Experimental Education 5, 1937, 296-312, b.

Schach, S.: The asymptotic distribution of the test statistic of the incidence test by Kannemann. - A correction. Biometrische Zeitschrift 18, 1976, 505-508.

Schneeweiß, H.: Entscheidungskriterien bei Risiko. Berlin: Springer, 1967.

Schulman, R.S.: A test theory model for ordinal measurements. Psychometrika 40, 1975, 455-472.

Schulman, R.S.: Correlation and prediction in ordinal test theory. Psychometrika 41, 1976, 329-340.

Schulman, R.S.: Individual distributions under ordinal measurement. Psychometrika 43, 1978, 19-29.

Schulman, R.S.: Ordinal data: An alternative distribution. Psychometrika 44, 1979, 3-20.

Scott, D., Suppes, P.: Foundational aspects of theories of measurement. The Journal of Symbolic Logic 23, 1958, 113-128.

Semenya, K.A., Koch, G.G.: Linear models analysis for rank functions of ordinal categorical data. Proceedings of the Statistical and Computational Section of the American Statistical Association 1979, 271-276.

Semenya, K.A., Koch, G.G., Stokes, M.E., Forthofer, R.N.: Linear models methods for some rank function analysis of ordinal categorical data. Communication in Statistics, Theory and Methods 12, 1983, 1277-1298.

Simon, G.: Alternative analysis for the singly-ordered contingency table. Journal of the American Statistical Association 69, 1974, 971-976.

Stevens, S.S.: On the theory of scales of measurement. Science 103, 1946, 677-680.

Suppes, P., Zinnes, I.L.: Basic measurement theory. In R.D. Luce et al. (Hrsg.): Handbook of mathematical psychology, Vol. 1, New York: Wiley, 1963.

Tallis, G.M., Dansie, B.R.: An alternative approach to the analysis of permutations. Applied Statistics 32, 1983, 110-114.

Thurstone, L.L.: A law of comparative judgement. Psychological Review 34, 1927, 273-286.

Thurstone, L.L., Chave, E.J.: The measurement of attitude. Chicago: University of Chicago Press, 1929.

Torgerson, W.S.: Theory and methods of scaling. New York: Wiley, 1958.

Upton, G.J.: The analysis of cross-tabulated data. New York: John Wiley & Sons, 1978.

Weitzenhoffer, A.: Mathematical structures and psychological measurements. Psychometrika 16, 1951, 387-406.

Williams, O.D., Grizzle, J.E.: Analysis of contingency tables having ordered response categories. Journal of the American Statistical Association 67, 1972, 55-63.

Williams, R., Zimmerman, D.: Reconsideration of the "Attenuation Paradox" - and some new paradoxes in test validity. Journal of Experimental Education 50, 1982, 164-171.

Yates, F.: The analysis of contingency tables with groupings based on quantitative characters. Biometrika 35, 1948, 176-181 u. corr. 424.

Yule, G.U.: On the association of attributes in statistics: with illustrations from the material of the childhood society. Philosophical Transactions of the Royal Society of London, Series A 194, 1900, 257-319.

Zimmerman, D., Williams, R.: The theory of test validity and correlated errors of measurement. Journal of Mathematical Psychology 16, 1977, 135-152.